Available online at www.sciencedirect.com

ScienceDirect

Poetics 41 (2013) 570–606

POETICS

www.elsevier.com/locate/poetic

Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding

Paul DiMaggio [a,*], Manish Nag [a], David Blei [b]

[a] Department of Sociology, Princeton University, 106 Wallace Hall, Princeton, NJ 08544, USA
[b] Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08544, USA

Available online 8 November 2013

Abstract

Topic modeling provides a valuable method for identifying the linguistic contexts that surround social institutions or policy domains. This article uses Latent Dirichlet Allocation (LDA) to analyze how one such policy domain, government assistance to artists and arts organizations, was framed in almost 8000 articles. These comprised all articles that referred to government support for the arts in the U.S. published in five U.S. newspapers between 1986 and 1997—a period during which such assistance, once noncontroversial, became a focus of contention. We illustrate the strengths of topic modeling as a means of analyzing large text corpora, discuss the proper choice of models and interpretation of model results, describe means of validating topic-model solutions, and demonstrate the use of topic models in combination with other statistical tools to estimate differences between newspapers in the prevalence of different frames. Throughout, we emphasize affinities between the topic-modeling approach and such central concepts in the study of culture as framing, polysemy, heteroglossia, and the relationality of meaning.

Keywords: Topic models; Polysemy; Heteroglossia; Meaning; Content analysis; National Endowment for the Arts

1. Introduction

This article addresses three puzzles. The first is methodological. How can researchers analyze large quantities of textual data efficiently and effectively? Specifically, how can we capture the information we need, reduce its complexity, and provide interpretations that are substantively

* Corresponding author.
E-mail addresses: dimaggio@princeton.edu (P. DiMaggio), mnag@princeton.edu (M. Nag), blei@cs.princeton.edu (D. Blei).

0304-422X/$ – see front matter
http://dx.doi.org/10.1016/j.poetic.2013.08.004

plausible and statistically validated? We present topic modeling and, specifically, Latent Dirichlet Allocation (LDA) as a promising solution to these challenges.

The second puzzle is theoretical: The sociology of culture has long been theory-rich and methods poor. Sociologists who study culture have generated numerous theoretical insights and developed concepts that promise a deep understanding of cultural change. Yet they have often lacked the means to make such concepts operational (Mohr, 1998). We suggest that topic modeling may provide a way to do just that for such central concepts as framing, polysemy, heteroglossia, and the relationality of meaning.

The third is part of an ongoing study by the first author of the dramatic politicization of government support for arts organizations and artists in the late 1980s after a two-decade honeymoon. This article addresses how press coverage of public funding of the arts evolved from 1986 to 1997, a period that spans the beginning and end of the period of most acute contention.

2. Background: the crisis of public arts support in the U.S.

U.S. municipalities supported museums as early as the nineteenth century and financed bands and orchestras into the 1920s; the Roosevelt administration created a federal jobs program for artists during the Great Depression; several states established arts councils before 1965, and arts organizations receive substantial tax subsidy through the charitable deduction.[1] Yet the United States had no permanent national program of grant support for the arts until President Lyndon Johnson signed legislation creating the National Endowment for the Arts (NEA) in September 1965. President Nixon expanded the NEA's budget dramatically and the agency used congressionally mandated block grants to build a network of state and, ultimately, local arts agencies, as well. Under the leadership of Chair Nancy Hanks, an associate of Nelson Rockefeller with strong ties to both Democratic and Republic legislators, the agency was popular and largely uncontroversial.

During the administration of Jimmy Carter, some Republicans criticized the Endowment for supporting what they alleged were socially relevant programs of little artistic merit. The NEA's real-dollar budget declined for the first time in 1980, due to a small nominal increase during a period of very high inflation. Nonetheless, the NEA's legislative support remained robust except among fiscal conservatives, one of whom, Ronald Reagan, was elected President in 1980. After a coalition of arts patrons and Republican insiders thwarted his initial plan to eliminate the NEA, Reagan appointed as Chair a long-time associate who effectively insulated the agency from serious conservative criticism during Reagan's two terms in office.

This changed with the inauguration of President George H.W. Bush in 1989. Beginning in May, congressional conservatives, backed by religious right movement organizations, vigorously criticized grants that supported works that critics found obscene (a Mapplethorpe retrospective that included homoerotic photographs) or sacrilegious (Andres Serrano's photograph of a crucifix immersed in urine, known as "Piss Christ"). Conservative groups, especially the American Family Association, began to focus attention on the NEA and to monitor actively its grants. Because the grant-making system gave much autonomy to program directors and review panels, and because many proposals (e.g., for juried exhibitions, fellowship competitions, or institutional support) did not describe specific work to be supported, it was easy for motivated critics to find potentially offensive artworks presented by organizations that had received federal money. It is not surprising that additional controversies ensued.

[1] The background paragraphs in Section 2 draw on the following sources: Alexander (2000), Burgess (2006), Dubin (1992), Fiss (1991), Frohnmayer (1993), Jensen (1995), Kimbis (1997), Koch (1998), Kresse (1991), and Ziegler, 1994).

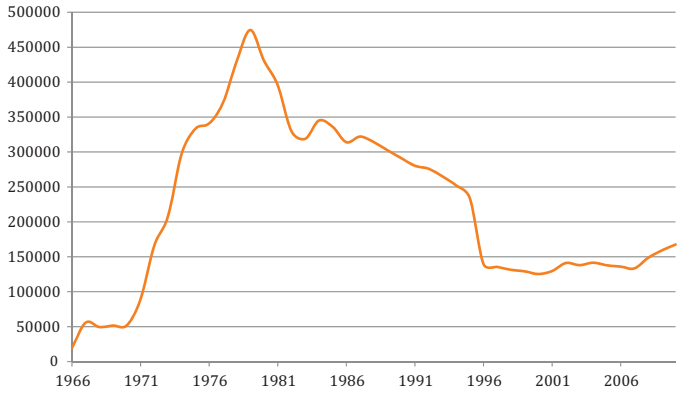

Fig. 1. NEA appropriations by year in thousands of 2010 dollars.
Source: National Endowment for the Arts. The 1976 Transition Quarter (extra quarter due to the transition to a new fiscal year ending in September rather than June) is omitted from this series. The authors adjusted the data using the CPI deflator.

Fig. 1 demonstrates the sharp growth in NEA appropriations (adjusted to 2010 constant dollars) through the late 1970s and precipitous declines between 1979 and 1983 and between 1995 and 1996. The first decline was driven more by inflation than by debates over the agency's grants; the second came after several years of political controversy, when Republicans took control of Congress in the 1994 midterm elections. The Endowment remained a target of Christian Conservatives (the Christian Coalition included the NEA's elimination as a promise in its "Contract with the American Family" in 1995), as well as fiscal conservatives. The Republican Congress tried and failed to eliminate the agency in 1995 and again in 1997. That year, President Bill Clinton appointed as Chairman Bill Ivey, a politically adroit folklorist with populist instincts, and appropriations stabilized. They rose slightly during the administration of George W. Bush, albeit to just over one third of their 1979 constant-dollar peak.

The reasons for the NEA's declining fortunes are not obvious. The NEA made potentially controversial grants from the beginning, but few were noticed; those that were, failed to produce enduring controversy before 1989. Conservative opposition to federal programs played a role, but the budget's constant-dollar decline began under President Carter, not President Reagan. Public opinion had not turned against the NEA before the controversy; polls showed that public support for federal arts funding increased during the 1980s, and declines lagged rather than preceded the political conflicts (DiMaggio and Pettit, 1999; Pettit and DiMaggio, 1997). Nor did the battles of the late 1980s reflect local community activism, according to studies of controversies over the arts and media in Philadelphia (where there were few controversies over grant-supported art exhibits or performances [DiMaggio et al., 2001]) and Atlanta (where heightened contention followed rather than preceded national conflicts [unpublished tabulations available on request]). If changes in public opinion and local activism cannot in themselves explain the crisis over arts funding in the U.S., then perhaps we can gain additional insight by examining the depiction of arts funding in the daily press.

3. Press accounts of public support for the arts in the U.S.

Did the tone of press coverage of government arts support change during the 1980s and 1990s? Did changes, if any, precede or coincide with political attacks on the NEA? Were some

newspapers more likely to frame arts support in a positive light, whereas others were more likely to frame it as offensive or contentious? In short, how did the press respond to, participate in or contribute to the NEA's political woes?

3.1. Why look at press coverage?

Why should we care about newspaper coverage of arts patronage? First, it provides clues as to what elites are thinking and doing. Newspapers cover topics when institutional actors, especially political figures, turn their attention to them, particularly when attention leads to extensive debate, legislative proposals, or executive action (Janssen et al., 2008; Molotch and Lester, 1974; Reese, 1991). Journalists are creative workers who read widely and interact intensely: Their writing reproduces representations current among public intellectuals. Moreover, news stories are often built directly around quotes (or paraphrases) from institutional actors (in the case of stories about public arts funding, artists, arts administrators, activists and politicians) and embody the assumptions and narratives those authorized speakers use to frame the topic at hand.

Second, press accounts are important because they influence the views of the reading public. Support for federal assistance to the arts eroded during the decade-long struggle over the NEA: The percentage of survey respondents supporting federal support for the arts, which had risen to 59 percent in 1987, declined modestly between 1987 and 1992 and fell to 45 percent by 1998. Similarly, the percentage of respondents favoring increased government aid to the arts declined slightly and the percentage wanting "much less" increased modestly (DiMaggio and Pettit, 1999; General Social Survey, 1998). Although one cannot establish a causal link, given that news reports represented the major source of information about the NEA for many Americans during this period, it seems likely that they contributed to this change.

But what of the decline in newspaper readership? The share of Americans reporting that they read a newspaper "every day" fell from 53 percent to 42 percent between 1986 and 1996; some people rarely read newspapers (13 percent in 1986 and 17 percent in 1996 reported reading newspapers never or less than once a week)[2]; and people who do, read selectively (Weaver and Mauro, 1978). Why then, aside from the fact that they are archived, should we focus on newspapers if we are interested in public views of government arts funding?

It is useful to rephrase this question in a more analytically precise manner: What are the mechanisms through which measurements of media content may tap factors that shape individual and collective perceptions and understandings? We believe that there are five:

(1) *Priming of existing schematic representations.* Among attentive readers who encounter press reports on a topic of interest, those reports may activate relatively well articulated schemata, strengthening existing positions (Iyengar and Kinder, 1987).[3] For arts supporters, positive news about NEA or state arts agency grants to praiseworthy or prestigious organizations and projects activates and reinforces prior views. Press coverage of government support for artworks deemed blasphemous or obscene does the same for cultural conservatives who are inclined to be suspicious of the arts (DiMaggio and Bryson, 2007). Coverage of budgetary conflicts may prime negative views of arts spending as wasteful among fiscal conservatives.

[2] Tabulations on General Social Survey website, annual frequencies for mnemonic NEWS ("How often do you read the newspaper—every day, a few times a week, once a week, less than once a week, or never?") http://www3.norc.org/GSS+Website/Browse+GSS+Variables/Mnemonic+Index/.

[3] Schemata are "knowledge structures that represent objects or events and provide default assumptions about their characteristics, relationships, and entailments under conditions of incomplete information" (DiMaggio, 1997, p. 269).

For such readers, both positive and negative arts-funding schemata may be easily activated, rendering the environment of representations especially influential. In the debates over the NEA, opponents may have been especially susceptible to mobilization, as they tended to hold more extreme positions than supporters. Moreover, between 1985 and 1990, opposition to the NEA, which was previously spread throughout the population, grew focused around party and church, as Republicans and Evangelicals became significantly more likely to oppose the agency (DiMaggio and Pettit, 1999).

(2) *Development of new representations.* Even readers with little or no interest in arts policy encounter references to government arts spending in articles devoted primarily to other topics, and, if repeated over time, such references may form new or clarify inchoate mental models, or they may render rarely activated associations chronically accessible (Price and Tewksbury, 1997). If one reads about an art exhibit one has enjoyed and notes that it was supported by an NEA grant, a positive association may form or be reinforced. Similarly, references to government funding in a lurid story about an exhibit with intense homoerotic or blasphemous images may engender or reinforce less positive associations.

(3) *Integration with broader schemata.* Inattentive readers with little interest in arts policy may integrate information about arts funding directly into broader social representations (political ideologies or politically relevant values or attitudes [Feldman, 2003]) with which press accounts articulate. For civil libertarians, news about restrictions on arts grants may prime, and nest easily within, broader concerns about bigotry and censorship. For cultural conservatives, news about government grants to controversial artworks might activate, and be incorporated into, more general narratives about government waste or moral decay.

(4) *Indirect influence through selective re-telling*. Social interactions play a critical role in the priming or formation of mental models. Readers or viewers discuss the day's news, often in the process communicating media representations to third parties (Bird, 2011). By providing the stuff of social talk, press accounts also provide opportunities for people with strong feelings to share their opinions in the form of vivid, memorable narratives. Accounts of controversial artworks may be for most people more salient, and more likely to be repeated, than references to an arts agency's financial support for an unexceptionable exhibit or performance.

(5) *Proxy value*. Even if we care only about citizens who do not read newspapers and whose friends do not talk about the stories in them, tracking press coverage may have value in so far as newspapers report what opinion leaders regard as important and newsworthy, and what newsmakers use multiple channels to disseminate (Boczkowski, 2010; McCombs and Shaw, 1972). Newspaper coverage of government patronage probably roughly tracked coverage in other media. Arts attenders who missed positive messages about federal and state grants in their local newspapers may have noticed them in brochures at art exhibits or in concert or theater programs. Religious conservatives who missed newspaper references to controversial grants may have encountered them in remarks from church pulpits or direct-mail appeals from conservative movement organizations.

In other words, press coverage both reflects and represents one stream of influence in the formation of elite and public opinion. If the press largely mentions arts agencies in connection with happy news about enjoyable or edifying exhibitions and performances, the agencies are likely to float along under a halo of good feeling. If the press more frequently mentions arts agencies in connection with troubling topics—sexual depravity, blasphemy, political conflict—that halo may turn into a cloud of negative associations. Superficially, at least, it appears that coverage of government arts funding took such a turn between 1986, before the dawn of the

Table 1
Newspapers used in study with number of qualifying texts in each.

Newspaper	Years covered	(*N*)
Houston Chronicle	1986–1997	1530
New York Times	1986–1997	1608
Seattle Times	1986–1997	2038
Wall Street Journal	1986–1997	323
Washington Post	1986–1997	2459

Total number of unique terms = 54,982; Total number of words (after exclusions) = 3,381,574.

"culture wars," and 1998, by which time Matthew Brenson (1998) wrote in the *New York Times* that "The National Endowment for the Arts has been described as embattled for so long that it now probably assumes this word is part of its name."

3.2. The data

We collected every article in the *Houston Chronicle*, the *New York Times*, the *Seattle Times*, the *Wall Street Journal*, and the *Washington Post* published between 1986 and 1997 that contained any reference to government support for the arts in the United States. We chose this period in order to have three years in the series (1986 through 1988) preceding the well-publicized Mapplethorpe controversy and to carry the series through 1997, the last year (at this writing) that a serious effort was mounted to eliminate the NEA. Criteria for choosing newspapers included (a) availability for automated search; (b) prominence; and (c) ideological and geographic diversity. The *New York Times, Wall Street Journal* and the *Washington Post* are major papers with national profiles, and in the period studied, they represented centrist, conservative, and liberal editorial views, respectively. The economic importance of the arts in New York means that the *New York Times* is in effect the hometown paper of much of the arts industry. The *Post* is the newspaper of record for matters involving Congress and the executive branch. Thus, both might be expected to have attended closely to news of the NEA. We included the Houston and Seattle papers to enhance regional diversity: Houston's visual arts institutions had experienced considerable growth in a relatively conservative social environment; and Seattle was notable for its active performing-arts scene and a liberal social milieu.[4]

Articles were identified by screening all records from 1986 to 1997 with a search algorithm that included three components: (1) explicit reference to the National Endowment for the Arts; (2) explicit reference to other arts agencies (mainly state and local); and (3) general references to public funding for the arts (rather than specific agency names).[5] Articles thus identified were

[4] We undertook pilot analyses on data collected in 2001 that included articles from *USA Today* (but omitting 1986) and *New York Newsday* (because other New York papers were not in digitally searchable data bases to which we had access at that time); and we were then only able to identify articles based on the headline and first paragraph of the *Houston Chronicle*, thus under-representing that paper. Topic-modeling results based on the pilot corpus were very similar to results based on the more complete corpus analyzed here, increasing our confidence in the robustness of these results.

[5] The search terms were (1) (*"national endowment" near3 arts) or ("arts endowment") or NEA*; (2): *(art$ near5 agenc$) or (art$ near5 council$) or (art$ near5 commission$) or ("arts funding") or (percent near2 art$)*; (3) *(art$ or opera or operas or theat$ or museum$ or orchestra$ or dance or exhibit or exhibition or gallery) near5 (government or federal or state or local or public) near5 (fund$ or assistance or money or aid or grant$ or support or contract)*). Dollar signs [$] are wild cards; numerals in "near" expressions indicate the distance [in number of words] within which associated words are sought.

screened manually to ensure that they fit inclusion criteria.[6] Table 1 includes a list of the newspapers and the number of valid articles discovered in each.[7]

Although the same algorithm was used to search each newspaper's digital archive, the number of articles varied. The *Wall Street Journal*, a national publication that emphasizes economic and financial news, printed far fewer stories mentioning government support for the arts than did the others, just 323 over eleven years. Coverage in the rest ranged from 1530 articles in the *Houston Chronicle* to 2459 in the *Washington Post*.

In designing this study, we made two critical choices. The first was to collect data on press attention to an *issue*—government assistance for the arts—rather than to *events* like the Mapplethorpe controversy or the introduction of legislation affecting the NEA (Rogers and Dearing, 1988). Doing so made it possible to track the relative prevalence of coverage focusing on contention as compared to stories that provided a more positive context.

A second consequential choice was to include virtually any article (including news stories, news analysis, and opinion pieces) that referred however marginally to tax-revenue-supported aid to the arts, rather than including only articles primarily devoted to that topic. We did so because we were interested not simply in the coverage of *conflicts*, but in change over time in the *environment of representations* surrounding government arts support, most of which is not controversial and much of which comes from state and local government. We included texts in which references to government arts support were marginal as well as central, because casual allusions may reflect prevailing assumptions better than carefully crafted reports and may be read by readers who would skip an article about government arts funding.

It is one thing to collect a corpus of almost 8000 texts and well over 3 million words. It is quite another to analyze it. To do so, we needed to reduce the complexity of the data in order to identify the principal themes that framed discussions of government support. In the next section, we describe how we used topic models for this purpose.

4. Topic modeling: an inductive relational approach to the study of culture

Textual analysis has always been a central part of the study of culture. The digitization of huge quantities of text has raised the stakes by enabling scholars to launch more ambitious projects, while requiring development of new, more powerful, analytic tools. As a leading text on content

[6] Topic modeling automates much of the culling process because, as noted below, it tends to quarantine text from inappropriate documents into particular topics. Hand-culling left open the option of using complementary analytic methods. The screening process excluded (a) articles that referred only to public arts funding outside the United States, to foundations or corporate or private patrons but not to government arts patronge, or to government grants for purposes other than support of arts programming or arts institutions (e.g., scientific exhibits at science museums); (b) articles about public museums that did not refer to funding; (c) articles about local cultural commissions that dealt solely with building licensing processes or about prizes that did not entail financial support; (d) articles that referred to an arts agency to identify someone by title (e.g., a list of persons attending a society wedding); (e) articles about the National Education Association (also initialized as NEA); and (f) articles with inapposite combinations of keywords (e.g., "Granted, the art of politics. . .").

[7] We believe that we succeeded in our effort to be comprehensive. A study that sought to collect articles on Mapplethorpe and the controversies over NEA funding of offensive art published from December 1988 through December 1989, using a range of methods, located 151 texts (McLeod and MacKenzie, 1998). By contrast, we identified 207 articles referring to Mapplethorpe alone in the 12 months of 1989, even though we examined only five sources and excluded articles about the photographer that did not refer to arts funding.

analysis puts it, digitization shifts "the bottleneck of content analysis from the costs of access and tedious human coding to the need for good theory, sound methodology, and software..." (Krippendorf, 2004, p. 43).

Sociologists ordinarily analyze texts in one of three ways. Some scholars simply read texts and produce virtuoso interpretations based on insights their readings produce. The limitations of this approach for generating reproducible results are apparent. The second common strategy is to produce a set of themes (based on research questions, theoretical priors, or perusal of a subset of texts), create a coding sheet, and then code texts by reading them (or, more often, by having research assistants read them) (Holsti, 1969). The limitations of this approach are (a) that it is impractical when corpora are very large; (b) the more analytically interesting are the research questions, the harder it is to achieve acceptable levels of inter-coder reliability; and (c) the approach presumes that the researcher knows what is worth finding in the texts before having analyzed them. A third strategy involves using computer programs to search texts for keywords (selected based on research questions or theoretical priors) and comparing subsets of texts with respect to the prevalence of those keywords (Stone et al., 1966). This approach requires the researcher to circumscribe the scope of exploration *a priori* and, by treating each instance of a term as equivalent, it violates a fundamental principle of cultural sociology, i.e., that meaning emerges from relations among terms rather than inhering within them. Although the latter two methods can be helpful in asking well defined questions of small sets of texts (and, as we demonstrate below, can complement inductive methods), neither is sufficient on its own for most analytic purposes.

It follows from this brief review that a sound approach to text analysis must satisfy four desiderata. First it must be *explicit*, so that data are available for the researcher to test his or her interpretations and for other researchers to reproduce the analyses. Second, it must be *automated*, in order to accommodate the volume of text available given the prevalence of digital archiving. Third, it must be *inductive* to permit researchers to discover the structure of the corpus *before* imposing their priors on the analysis, and to enable different researchers to use the same corpus to pursue different research questions. Finally, it must *recognize the relationality of meaning* by treating terms as varying in meaning across different contexts. Topic modeling (Blei, 2011; Blei and Lafferty, 2009) satisfies all four conditions.

4.1. How topic models work

Topic modeling algorithms are a suite of machine learning methods for discovering hidden thematic structure in large collections of documents. With a collection of documents as input, a topic model can produce a set of interpretable "topics" (i.e., groups of words that are associated under a single theme) and assess the strength with which each document exhibits those topics. Topic models enable researchers to code text collections that are too large to code by hand—a topic model will estimate a coding instrument and situate each document within it. Furthermore, a topic model might uncover topics that a researcher using hand coding methods might not otherwise have seen. For researchers in the social sciences, topic models provide a new computational lens into the structure of a collection of texts. With topic models, researchers can discover new patterns in their text data and analyze much larger collections than is possible by hand.

In this work we used Latent Dirichlet Allocation (LDA), which is among the simplest topic models (Blei, 2012; Blei et al., 2003). LDA is a statistical model of language. It assumes that

there are a set of topics in a collection (the number is specified in advance), where a topic is formally defined as a distribution over a vocabulary. Terms that are prominent within a topic are those that tend to occur in documents together more frequently than one would expect by chance. In LDA, each document exhibits those topics with different proportions. For example, the model described below contains topics about ''controversial NEA grants,'' ''performing arts,'' and ''urban arts projects.'' (By a topic being ''about'' a subject, we mean that those distributions over the vocabulary place high probability on words that an analyst would interpret as related to the subject.) Articles that discuss controversial NEA-supported art exhibits and dance, for example, will exhibit the first two topics, respectively; articles that discuss city policies related to the arts will exhibit the third; and an article about NEA support for controversial performing-arts events will exhibit the first and the second. We emphasize that these topics are not known in advance. The algorithm behind LDA analyzes the collection to estimate simultaneously the topics and how the documents exhibit them. Table 2 illustrates the most frequent words from topics uncovered in our collection of articles relating to government support for the arts.

LDA takes a relational approach to meaning, in the sense that co-occurrences are important in the assignment of words to topics. Intuitively, in order to capture these patterns of co-occurrence, LDA trades off two goals: first, for each document, allocate its observed words to few topics; second, for each topic, assign high probability to few words from the vocabulary. Notice that these goals are at odds. Consider a document that exhibits one topic. Its observed words must all have probability under that topic, making it harder to give few words high probability. Now consider a set of topics, each of which has very few words with high probability; documents must be allocated to several topics to explain those observations, making it harder to assign documents to few topics. LDA finds good topics by trading off these goals.

We have described the intuitions behind LDA. The algorithms for LDA, however, are derived by taking a Bayesian probabilistic perspective (Gelman et al., 2003)—encoding the topics and the per-document topic proportions as hidden random variables in a hierarchical probabilistic model and then approximating the conditional distribution of those variables given an observed collection of documents. In this article, we analyze the output of such algorithms (Blei et al., 2003; Griffiths and Steyvers, 2004). Given the texts, LDA inference produces a set of topics (see Table 2), and for each document, an estimate of its topic proportions and to which topic each observed word is assigned.

For sociologists of culture, an important element in topic models' appeal is the interpretability of some or all of the topics in most solutions. Such substantive interpretability is *not* required by the approach (Blei et al., 2003, p. 996), as for many purposes (e.g., identifying documents at risk of containing information requiring redaction) interpretation is unnecessary. Moreover, the program ''knows'' only where each text begins and ends and what terms are contained within it (with no semantic information about the terms themselves). Blei (2012, p. 79) attributes the interpretability of most topics to ''the statistical structure of observed language and how it interacts with the specific probabilistic assumptions of LDA.'' In applications to the study of culture, substantive interpretability is crucial. Many topics may be viewed as *frames* (semantic contexts that prime particular associations or interpretations of a phenomenon in a reader) and employed accordingly (Gamson, 1992; Klebanov et al., 2008).

Another particular strength of topic modeling is its ability to capture polysemy and disambiguate different uses of a term, based on the context (other terms) in which it appears. In their emphasis on *relationality*, topic models capture the insight, shared by linguistics and much cultural sociology, that meanings emerge out of relations rather than residing within words

Table 2
12-Topic solution, unsupervised topic model, 100 Highest-Ranked Terms Per Topic [dark shading (2, 5, 8) = Conflict Topics; Light Shading (1&7) = local government projects and funding; no shading (3, 4, 6, 9, 10, 11, 12) = specific arts genres, events, or grant purposes; Alpha was set to .1, Eta to .08, and the program ran through 50 iterations].

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
city	nea	music	tv	senate	arts	budget	bush	film	theater	information	art
building	art	orchestra	film	house	organizations	tax	government	book	dance	festival	museum
park	endowment	jazz	show	budget	museum	percent	political	poetry	company	saturday	artists
design	frohnmayer	symphony	television	congress	groups	county	president	children	ballet	tickets	gallery
downtown	arts	opera	news	bill	artists	council	clinton	black	theatre	sunday	paintings
art	artists	concert	channel	clinton	school	city	campaign	writing	play	22	exhibition
project	mapplethorpe	musicians	global	republicans	art	money	buchanan	writers	broadway	call	artist
center	helms	concerts	http	appropriations	grants	state	right	poet	production	center	show
commission	grants	musical	icon	rep	council	board	republican	story	season	children	painting
public	funding	band	movie	federal	00	government	abortion	mother	festival	free	collection
sculpture	grant	composer	night	committee	center	federal	helms	school	theaters	park	work
street	agency	blues	president	r	director	cuts	conservative	ms	artistic	730	works
murals	congress	hall	war	republican	education	increase	party	literary	performance	noon	sculpture
buildings	obscenity	classical	cbs	vote	community	plan	democratic	women	actors	county	exhibit
space	chairman	piano	hollywood	spending	students	income	issue	books	dancers	admission	century
site	artistic	composers	star	funding	board	00	election	father	plays	friday	sculptures
square	obscene	orchestras	pbs	gingrich	program	cut	issues	read	director	21	objects
area	council	pianist	class	amendment	support	taxes	public	films	productions	nw	painted
mural	corcoran	singers	york	congressional	cultural	services	politics	novel	opera	show	images
foot	federal	conductor	texas	agency	fund	proposed	speech	poems	stage	street	museums
construction	censorship	songs	air	cuts	symphony	programs	rights	writer	musical	oct	curator
built	decision	quartet	los	gop	grant	officials	americans	didn	performances	sept	contemporary
town	serrano	festival	police	chairman	commission	spending	religious	stories	companies	tomorrow	view
avenue	yesterday	mozart	nbc	sen	schools	mayor	reagan	write	audience	avenue	pieces
architect	amendment	season	station	subcommittee	programs	funds	candidate	love	shakespeare	arts	painter
community	panel	chamber	radio	legislation	funding	deficit	voters	work	repertory	commission	galleries
artists	controversial	performances	abc	president	money	development	democrats	don	troupe	aug	pictures
county	controversy	performance	actor	agencies	institutions	public	liberal	fiction	playwright	church	display
neighborhood	political	philharmonic	soviet	cut	foundation	economic	policy	kids	ms	sponsored	abstract
wall	artist	song	documentary	democrats	performing	funding	conservatives	movie	playwrights	library	glass
hall	exhibit	singer	angeles	fiscal	opera	housing	race	reading	center	series	modern

Table 2 (*Continued*)

architects	show	singing	video	endowment	endowment	pay	america	poets	actor	thursday	photographs
artist	radice	folk	world	administration	funds	commission	culture	young	cast	jersey	drawings
historic	members	play	oscar	y	theater	district	christian	literature	show	annual	catalogue
residents	national_endowme	album	producer	humanities	city	fiscal	republicans	family	arena	film	portrait
architecture	nt_arts	playing	hour	nea	president	property	social	published	choreographer	july	color
museum	gallery	ensemble	network	percent	executive	cost	dole	friends	audiences	wednesday	exhibitions
feet	support	perform	wife	voted	museums	business	candidates	lives	the	nov	paint
metro	finley	musician	vietnam	speaker	county	revenue	values	child	premiere	concert	the
projects	expression	rock	movies	yesterday	organization	department	congress	author	drama	school	japanese
station	jesse	beethoven	yesterday	programs	cach	health	tax	parents	performed	college	american
spaces	debate	recording	films	approved	humanities	proposal	gay	aids	dancer	dc	portraits
garden	photographs	audience	stations	interior	percent	hotel	anti	right	theatrical	31	ms
parking	aids	guitar	broadcast	government	budget	costs	american	teacher	kennedy	saturdays	wood
portland	performance	soprano	tonight	armey	chairman	financial	presidential	voice	performing	tuesday	visual
wpa	government	sing	actress	democrat	awards	fund	white	thought	rep	gallery	landscape
walls	money	dance	celebrity	n	corporate	program	moral	students	work	73	world
plan	offensive	chorus	magazine	lawmakers	national	committee	administration	real	directors	program	piece
designed	court	violin	won	neh	visual	approved	coalition	son	american	art	smithsonian
facility	freedom	gospel	illustration	plan	local	sales	welfare	poem	tony	event	installation
statue	decency	instruments	game	national	projects	police	platform	living	choreographers	monday	collections
place	issue	concerto	stars	proposed	project	companies	women	writes	i	sundays	furniture
artwork	letter	american	dinner	education	dance	employees	state	written	directed	museum	materials
bridge	restrictions	series	story	leaders	raising	businesses	civil	wrote	musicals	theater	forms
urban	sexual	played	bill	votes	teachers	insurance	war	men	black	students	figures
parks	funds	choir	party	senators	children	reduce	agenda	self	gockley	include	great
south	funded	performed	car	funds	committee	schools	economic	age	hgo	feb	collectors
00	public	program	named	endowments	public	support	politicians	characters	rehearsal	council	photography
north	work	violinist	vice	members	awarded	office	primary	boy	ensemble	workshop	shows
kent	alexander	sound	business	x	nea	agencies	taxes	girl	joffrey	march	wall
land	statement	string	broadcasting	leadership	state	private	society	kind	act	930	light
near	director	players	rock	voting	announced	school	education	sense	choreography	830	sculptor
streets	panels	schwarz	entertainment	senator	financial	construction	senator	learn	night	road	retrospective
glass	wildmon	bach	murder	yates	culture	service	nation	talk	graham	adults	19th
plaza	congressional	ellington	lady	defense	received	members	gop	friend	studio	music	picasso
visitors	sen	recital	games	medicare	members	care	north	culture	york	lecture	prints
piece	hughes	recordings	awards	floor	educational	growth	support	come	playhouse	open	walls
installed	language	carnegie	o	williams	nonprofit	revenues	federal	press	dancing	dance	corcoran
location	review	instrument	media	legislative	staff	management	gantt	started	staged	society	de

district	community	premiere	sports	veto	contributions	residents	reform	teaching	comedy	25	paris
architectural	museum	bernstein	cc	newt	university	aid	jesse	language	shows	dec	colors
open	legislation	conducting	internet	measure	major	jobs	convention	eyes	tickets	26	painters
river	andres	performing	guy	democratic	donors	arts	debate	american	de	crafts	history
plans	peer	audiences	industry	majority	orchestra	law	care	world	touring	28	black
renovation	exhibition	performers	cable	debate	ms	increases	power	born	producer	area	french
floor	pornography	pop	press	department	group	workers	views	prize	performers	tonight	whitney
wright	gay	traditional	award	eliminate	ballet	agency	constitutional	taught	actress	featuring	studio
seat	contemporary	player	market	conservatives	artistic	administration	leaders	remember	perform	april	seen
island	protest	evening	national_endowment_arts	conference	music	economy	liberals	room	balanchine	events	berlin
development	explicit	bands	food	passed	american	dollars	carolina	daughter	alley	reservations	sense
developers	york	orchestral	screen	conservative	annual	executive	argument	recalls	papp	scheduled	bronze
cost	board	works	commercial	white	private	raise	georgia	sitting	acting	jan	canvas
places	reauthorization	radio	singer	national_endowment_arts	development	voters	family	wife	works	hall	organized
local	judge	solo	michael	compromise	society	community	country	death	ballets	king	form
memorial	law	cello	simpson	proposal	award	motel	sex	character	ticket	exhibit	curators
northwest	applications	bass	british	tax	association	rate	view	brother	morris	feature	craft
landscape	ne	choral	baseball	bills	national_endowment_arts	needs	black	husband	modern	fridays	van
water	urine	young	celebrities	reduction	professional	spend	democracy	college	costumes	community	garden
house	rejected	tenor	john	gorton	college	groups	reed	play	summer	27	20th
build	advisory	guitarist	top	increase	report	eliminate	polls	english	albee	630	metal
library	approved	repertoire	disney	appropriation	elementary	projects	television	felt	tap	women	design
commissioned	national	juilliard	hit	capitol	study	fees	groups	learned	ailey	university	culture
everett	process	soloist	club	alexander	receive	legislature	believe	sex	martha	29	fine
trees	meeting	trio	host	reagan	foundations	meeting	free	great	presented	connecticut	landscapes
entrance	orr	cellist	joe	bush	district	higher	crime	body	pnb	food	chinese
mall	organizations	tickets	moscow	secretary	texas	capital	pro	doesn	dances	workshops	paper
brick	case	quintet	kennedy	hackney	donations	facilities	term	library	stages	reception	human
spafford	cheney	piece	united	representative	dc	local	united	hair	producers	24	different
home	standards	culture	france	regula	corporations	increased	majority	stuff	graney	town	auction
sites	cahall	flute	shot	delay	activities	total	elected	grew	city	ages	hirshhorn
	senator										

(Mohr, 1994; Mohr and Duquenne, 1997; Saussure, 1983). Thus many terms may appear in more than one topic within a given corpus of documents.

A third virtue of topic modeling is its deep affinity to the central insight in the sociology of culture that texts do not necessarily reflect a singular perspective but are often characterized by *heteroglossia*, the copresence of competing "voices"—perspectives or styles of expression—within a single text (Bakhtin, 1982 [1934–1941]).[8] Blei (2012, p. 78) writes that the fundamental "intuition behind LDA is that documents exhibit multiple topics." The results that LDA produces can be useful in examining heteroglossia empirically.

4.2. Implementation of the model on the arts-funding corpus

Topic modeling is an exploratory technique, useful for imposing order upon large bodies of textual data and for discovering information that helps analysts see beyond their priors. Topic models can produce any number of topics that the researcher specifies: One chooses the number based on interpretability and analytic utility (Blei and Lafferty, 2009, p. 12). Like any clustering technique, the method should be employed as a heuristic tool in combination with additional information by a research team that includes subject-area experts (Grimmer and Stewart, 2011).

In the analyses reported here, each text is a newspaper article that mentions government support for the arts. The analysis was conducted on all 54,982 terms appearing in all 7958 documents—save for names of newspapers or their cities of publication (which would have artificially biased the solution toward topics that treated articles from the same newspaper as similar), and "stop words" (very common words like articles, conjunctions, or forms of the verb "to be"). We included all texts in one analysis, thus assuming that a single underlying structure characterized discourse about government arts support in all five newspapers. This enabled us to examine variation in newspapers' relative emphasis on particular themes, at the expense of investigating variation in the topic structure across sources.

Think of the model as a lens for viewing a corpus of documents. Finding 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. Just as different lenses may be more appropriate for long-distance or middle-range vision, different models may be more appropriate depending on the analyst's substantive focus. As the statistician George Box (1979, p. 202) wrote of models that cluster population data in order to make them tractable: "All are wrong; some are useful."

In some cases (for example, identifying public records that require redaction), the relative efficacy of different models can be readily assessed. But when topic modeling is used to identify themes and assist in interpretation, rather than to predict a knowable state or quantity, there is no statistical test for the optimal number of topics or for the quality of a solution. Indeed, a statistical test for an overall solution (as opposed to for the quality of particular topics, for which assessment methods exist [Mimno and Blei, 2011]) would be misleading, because models often shunt noisy data into uninterpretable topics in ways that strengthen the coherence of topics that remain. Thus, the test of the model as a whole is its ability to identify a number of substantively meaningful and

[8] Applications of topic modeling of particular interest to social scientists (Ramage et al., 2009) include analyze of over-time change in and influence among scientific texts (Blei and Lafferty, 2009; Gerrish and Blei, 2010); discovery of groups within and networks among S&P 500 corporations (Doyle and Elkan, n.d.); predicting congressional roll-call votes (Gerrish and Blei, 2011), political agendas (Grimmer, 2010) and legislative issue salience (Quinn et al., 2010); and ranking universities based on the extent to which their research is cutting edge (Ramage et al., 2010).

analytically useful topics, not its success in optimizing across all topics. Appendix A in the online supplement describes the process by which we selected the model on which the analyses that follow are based.

4.3. Interpreting the solution

Table 2 displays the 12-topic solution, listing the 100 highest-ranked terms for each topic (based on TF*IDF weighting, which adjusts prevalence within a topic for prevalence within the corpus as a whole). Topic order has no significance. We call attention to three sets of topics: Three topics highlighting different types of social or political conflict (dark shading); two topics emphasizing local projects and revenues (light shadings); and seven topics primarily concerned with specific arts genres, types of grant, or event information (no shading). To guide the reader through the process of interpreting model results, we begin our discussion with an extended account of the conflict topics (2, 5, and 8), and then discuss the others more succinctly.

4.3.1. Conflict topics

These topics are the focus of our analysis of press coverage of debates over the National Endowment for the Arts. Three topics define different conflict-laden semantic contexts. The first of these, Topic 2 comprises terms related to controversies over NEA grants (see Table 2). We begin our interpretation by visually inspecting the terms, which are ranked on the basis of their centrality to this topic (i.e., the extent to which they appear with other terms in the topic, adjusted for their prevalence in the corpus). The first terms make it clear that this topic pertains to the National Endowment for the Arts: *NEA, art, endowment, Frohnmayer* (NEA Chair during the most tumultuous period), *arts*, and *artists*. All of these terms refer to the agency, its leadership, or the objects of its benefactions. Next we come to terms directly related to the controversies that bedeviled the agency in the late 1980s: *Mapplethorpe*, *Helms* (Sen. Jesse Helms, the NEA's leading congressional antagonist), *grants*, *funding*, *agency*, and *congress*. Other terms reinforce this theme: *chairman*, *obscene*, *corcoran* (the Corcoran Museum, which canceled the Mapplethorpe exhibit, an act condemned by much of the art world), *censorship*, *Serrano*, *panel*, *controversial*, *controversy*, and names of several other artists and political figures central to the debates.

After reviewing the list of terms, the next step in interpretation is to examine texts that exhibit Topic 2 with the highest probability. Recall that each word in each text (stop words excepted) is assigned to one of the twelve topics. We look now at articles in which the highest proportions of words are assigned to Topic 2.

The five articles in which this topic accounted for all assigned terms included three covering a lawsuit by artists who claimed that the NEA exercised censorship in denying them grants after its panels had selected their projects for funding, and a news article and an editorial about the NEA's decision to withhold funding from an art exhibit on AIDS after learning that the exhibit catalog criticized prominent political and religious figures. Titles included "National Endowment for the Arts to Settle Suit for $252,000" (*Houston Chronicle*, June 6, 1993); "NEA to Pay 4 Denied Art Grants, but 'Decency' Rule Challenge Unresolved" (*Washington Post*, June 5, 1993); "NEA Withholds Grant for AIDS Art Exhibit" (*Houston Chronicle*, November 9, 1989); "Cardinal Principle" (*New York Times*, November 22, 1989 [editorial deploring the NEA's decision on the AIDS exhibit]); and "Arts Grant Decency Standards Struck Down" (*Seattle Times*, June 9, 1992, describing a federal court decision that was subsequently appealed). Inspection of these articles, as well as the next twenty-five ranked by the percentage of words assigned to Topic 2, all of which were squarely on topic, confirmed our interpretation that the topic focused on controversial

grants, objections to those grants, the NEA's efforts to appease its critics, and artists' reactions to those efforts.

Topic 5 includes terms related to congressional deliberations and the politics that accompanied them (see Table 2). Congress became involved in the NEA controversies in three ways: repeatedly postponing the agency's reauthorization; cutting or threatening to cut its budget; and proposing or enacting limits to the agency's autonomy. The most highly ranked terms are: *senate, house, budget, congress, bill, clinton, republicans, appropriations, rep* (as in "Representative"), and *federal*. Words specifically related to cultural funding (*NEA, NEH, Alexander* [Jane Alexander, Pres. Clinton's first NEA Chair], and *Interior* [the House Committee that oversees the cultural agencies' budgets]) also appear on the list, indicating that cultural agencies are mentioned in the context of broader budget issues.

Indeed, texts in which Topic 5 terms dominated word assignments dealt with congressional actions related to the NEA's budget. Headlines of articles in which more than 97 percent of assigned words were assigned to Topic 5 included "Votes in Congress" (*New York Times*, October 17, 1993, noting a negative vote on abolishing the NEA and a positive vote reauthorizing the agency), "For the Record" (*Washington Post*, September 26, 1991, discussing Senate votes—including votes against cutting the NEA budget and for imposing restrictions forbidding grants in support of obscene artworks); "How Texans Voted" (*Houston Chronicle*, July 18, 1993, reporting votes on a motion to defund the NEA); and two similar articles in the *Chronicle* reporting Texas legislators' votes on other matters affecting the Arts Endowment. In addition to such brief informational items, other articles in which Topic 5 terms were prevalent bore such titles as "NEA Funding Dealt Blow in House" (*Houston Chronicle*, July 14, 1995), "Future of National Endowment For Arts Looks Uncertain as Legislation is Stalled" (*Seattle Times*, July 25, 1990), and "Senate Panel Backs Money for Arts Agency" (*New York Times*, July 19, 1997).

Topic 8 is an important one for this study, as it represents the integration of conflict over the arts into a broader frame of social and political conflict associated with the "culture wars" of the 1990s (see Table 2). Whereas Topic 2 contexts depict the NEA's controversial grants as isolated problems, Topic 8 contexts tend to portray arts-funding controversies as one instance of a broader cultural struggle, marking their transformation from a series of *events* to an *issue* aligned with other moral or social issues (Shaw, 1977). The highest ranked terms in this topic are related to electoral politics (*Bush*, *political*, *president*, *Clinton*, *campaign*, and *Buchanan*), reflecting the influence of Pat Buchanan, who called for a "culture war" during his challenge to the first President Bush for the 1992 Republican presidential nomination. Other terms on the list refer to ideological and cultural issues directly (e.g., *culture*, *abortion*, *gay*, *Christian*, *moral*, *sex*, *religious*, and *family*).

News stories tend to be episodic rather than thematic (Iyengar, 1991), event-driven rather than analytic, and inattentive to the broader context in which events occur. Thus if our interpretation of Topic 8 is correct, we would expect that, in contrast to the top-ranked articles for Topics 2 and Topic 5, which were all news reports, the top-rated articles for Topic 8 would include news analyses, editorials, and op-eds—i.e., genres more likely than news stories to place events in context. Indeed, this was the case. Articles in which words were most likely to be assigned to Topic 8 (more than 90 percent in each case) included "Bush's Message Might Be: I Like the Job" (*Houston Chronicle*, March 7, 1992, news analysis of the Republican primary battle between Bush and Buchanan); "The Republican Platform: Excerpts From the Republican Party's Platform: A New Call for Unity" (*New York Times*, August 18, 1992); "Bush's Spent Presidency" (*Washington Post*, March 6, 1992, editorial lamenting Bush's responsiveness to the right, citing Bush's firing of the NEA Chairman as an example); "Bookshelf: Middle Class Left in the Lurch" (*Wall Street Journal*, May 31, 1991, a review of E.J. Dionne's *Why Americans Hate*

Politics that referred to ''federal grants for obscene art''); and ''God and the GOP; Will We on the Christian Right Go Wrong?'' (*Washington Post*, September 26, 1993, an op-ed column by the head of the Christian Action Network citing the campaign against NEA as one of several cases in which the Christian Coalition had lost its focus. Note that none of these pieces was a regular news story: Topic 8's natural home was in longer analysis pieces that characterized broadly the state of American politics or culture.[9]

4.3.2. Urban topics

Two topics deal with the urban environment. Topic 1 includes numerous terms related to the role of the arts in economic-development schemes and the use of art to enhance urban places: *city*, *building*, *park*, *design*, *downtown*, *art*, *project*, *center*, *commission*, *architecture*, *murals*, *historic*. Articles in which this topic was strongly represented included stories about a 25-mile bike trail featuring artistic installations and about the renovation of a monument in a Houston park. Articles in which Topic 1 was dominant tended to focus on local arts agencies rather than state or federal arts support. Topic 7 is a state/local counterpart to Topic 5, dominated by terms related to state and local budgets and financial deliberations like *budget*, *tax*, *percent*, *county*, *council*, *city*, *money*, *state*, *board*, *government*, *cuts* and *services*. Almost all articles in which Topic 7 was prevalent described tax or budget issues, including several on Seattle's and Houston's use of hotel/motel-tax revenues to assist cultural organizations.

4.3.3. Genre topics

With one exception, these topics refer to particular art forms or types of grant. Many articles in which these topics are prominent focus on the content of artistic exhibits or performances that received government grants. Topic 3 covers all kinds of musical performances and organizations, with high-culture forms (*orchestra*, *jazz*, *symphony*, *opera*) ranked highest, but more popular forms (*band*, *blues*, *folk* and *rock*) included as well. The articles that Topic 3 terms dominated most thoroughly previewed a chamber music festival, described a chamber-music concert series, and profiled a jazz composers' orchestra. Topic 10 plays a similar role for theater and dance: Top-ranked terms include *theater*, *dance*, *company*, *ballet*, *theater*, *play*, *Broadway*, *production*, *season*, and *festival*. Topic 10 terms were featured most prominently in articles describing the opening of new theater and dance productions supported by government grants. Finally, Topic 12 includes terms referring to museum exhibits and the visual arts: *art*, *museum*, *artists*, *gallery*, *paintings*, *exhibition*, *artist*, *show*, *painting*, *collection*, and *works*. The esthetic emphasis of this topic is clear from such terms as *forms*, *figures*, *light*, *colors*, and *sense*. The articles in which Topic 12 was most heavily represented included accounts of exhibits of antiquities and of modernist paintings.

Two topics within the genre subset dealt with the media and works often presented by commercial rather than nonprofit entities. Topic 4 consists of terms primarily associated with television or film production, such as *tv, film*, *show*, *television*, *news*, *channel*, *movie*, *cbs*, *pbs*, *documentary*. Topic 4 is represented in articles referring to grants to documentary filmmakers. Other articles, including several in which the topic is most prominent, refer to arts funding in passing. One top-rated article, for example, mentioned that an Academy Award winner wore a lapel button symbolizing political support for the NEA. Topic 9 also includes *film* and *movie* among its ranked terms, but along with terms related to narrative (including biographical sketches or plot summaries) and the literary arts: *book*, *poetry*, *children*, *black*, *writing*, *writers*,

[9] Analyses usin the STATA fmlogit procedure with controls for year and source confirmed the association between article length and the prevalence of Topic 8. (Results available upon request.)

poet, *story*, and *mother*. The article in which Topic 9 was most prevalent was a profile of Sapphire, the author of the novel *Push*, on which the prize-winning film ''Precious'' was based, who had received a poetry fellowship from the NEA.

The final two topics in this subset belong in this group for other reasons. Topic 6 focuses upon grants to many kinds of arts organizations, often for education or community-outreach. It combines terms related to arts support (*arts*, *organizations*, *museum*, *grants*, *council*, *commission*) with terms related to education (*school*, *education*, *community*, *students*, *board*). Topic 11 comprises terms that appear in announcements of events, like *information*, *festival*, *Saturday*, *tickets*, *Sunday*, *call*, *children*, *free*, *park*, and *7:30*. Topic 11 terms focus on where and when events take place and how much they cost, and are represented most heavily in event listings. The topic is of no substantive interest, but its emergence sharpens genre-related topics by segregating boilerplate terms in a distinct location.

4.4. Working with the solution

Producing an interpretable solution is the beginning, not the end, of an analysis. The solution constructs meaningful categories and generates corpus-level measures (e.g., the percentage of documents in which a given topic is highly represented) and document-level measures (e.g., the percentage of words in each document assigned to each topic) based on these categories. It remains for the analyst to use this information to address the analytic questions that motivated the research. The analyst must also validate the solution by demonstrating that the model is sound and that his or her interpretation is plausible.

There are three forms of validation. The first is *statistical*, seeing if the model results are consistent with the assumptions of the model and, if they are not (as is usually the case), using deviations to better interpret the results. The use of the mutual information (MI) criterion (Mimno and Blei, 2011) represents a test of this kind (see Section 6.1). The second is *semantic* or *internal* (Grimmer and Stewart, 2011): For this, we employ hand coding of sample texts to discover whether the model meaningfully discriminates between different senses of the same or similar terms (see Section 5.1). The third is *predictive* or *external* (Grimmer and Stewart, 2011): Here we ask if attention to particular topics responds in predictable ways to news events that should affect their prevalence if our interpretations are correct (see Section 5.3). We explore these forms of validity below, in the context of a broader discussion of the use of topic models for operationalizing key concepts in the sociology of culture.

5. Topic modeling renders operational central ideas in the sociology of culture

Although automated approaches to textual analysis are increasingly plentiful, specific affinities between the topic-modeling algorithm and key ideas in the sociology of culture produce a strong fit between theory and method. In this section, we discuss the ways in which topic models enable scholars to render operational ideas about the relationality of meaning, heteroglossia, and framing.

5.1. The relationality of meaning, contextual polysemy, and semantic validation of the model

It is axiomatic to most social-scientific approaches that meaning is relational—i.e., that meanings do not inhere in symbols (words, icons, gestures) but that symbols derive their meaning

from the other symbols with which they appear and interact (Mohr, 1994; Mohr and Duquenne, 1997; Saussure, 1983). This assumption is built into the DNA of topic models in two ways. First, the LDA algorithm allocates terms to topics based on their relations to other terms, placing together terms that appear in the same texts more frequently than one would expect by chance. Second, the algorithm permits different instances of the same term (i.e., different words) to be assigned to more than one topic, in effect treating the term as a different unit of meaning depending on the semantic environment in which it is instantiated. Thus, LDA provides a means of operationalizing contextual *polysemy*, in which certain terms take on different meanings depending upon the context in which they appear (Copestake and Briscoe, 1995).[10] In LDA, each topic can be viewed as a distinct discursive context for a term that is sometimes assigned to it. In the arts-funding corpus there are many examples of terms that appear in the top 100 terms of several topics—e.g., film (6 topics), museum (6 topics), council (7 topics), park (7 topics), music (7 topics), moral (5 topics), and many others.

This feature enhances the realism of topic-model solutions and their ability to capture major themes in corpora of texts. It also provides a means for *semantic validation* of topic-model solutions. If the algorithm works properly, when the same term is assigned to different topics, different meanings should be evident. And if the analyst has interpreted the topics correctly, these differences should be consistent with the analyst's interpretation. In effect, then, examining differences in the meaning of the same term when it is assigned to different topics serves as a test of internal validity.

We illustrate this point with an analysis of the term "museum," which appears in 3271 documents and is assigned to six different topics. We compare the senses of the term when it is assigned, respectively, to Topic 1 (urban space), Topic 6 (grants to multiple organizations and multiculturalism), and Topic 12 (exhibitions), the three topics in which it is most prevalent.[11] We identified every text in which "museum" was assigned to a given topic and categorized the texts into five sets according to the prevalence of the topic (i.e., the percentage of words assigned to it in the text): 90 percent or more; 65–89 percent; 35–64 percent; 10–34 percent; and less than 10 percent. Except for those cases in which there were too few texts in a set, we randomly sampled 12 texts (for each topic in each set). These texts were then coded by hand, using the first occurrence of the term assigned to a given topic in each text.

Results are displayed in Fig. 2. The *y* axis represents a count of texts falling into each category. The top panel displays results for the full sample, the middle panel for texts with 35 percent or more of words assigned to the focal topic, and the bottom panel for texts with fewer than 35 percent of words assigned to the focal topic. "Core" refers to assignments to contexts that are consistent with our interpretation of each topic.

Because we interpreted Topic 1 as associated with the built environment, public art, and economic development, its core subjects include references to museums in the context of urban development plans, new museum buildings or major renovations of old ones, and public art works. Because we interpreted Topic 6 as referring to both grants to multiple art forms and arts outreach, we define its core subject areas as reports of major grants or gifts to museums and references to multiculturalism, museum education programs, community outreach, and the

[10] The terms *polyseme* and *homonym* are often used as synonyms, but we follow conventional usage in semantics by distinguishing *polysemes* (terms with common origins but distinct, albeit related, senses) from *homonyms* (etymologically unrelated terms with the same spelling but different meanings) (Copestake and Briscoe, 1995, p. 15).

[11] In Appendix B of the online supplement, we provide a second example, analyzing the appearance of the term "film" in Topic 4 (media) and Topic 9 (narrative) using the same procedure.

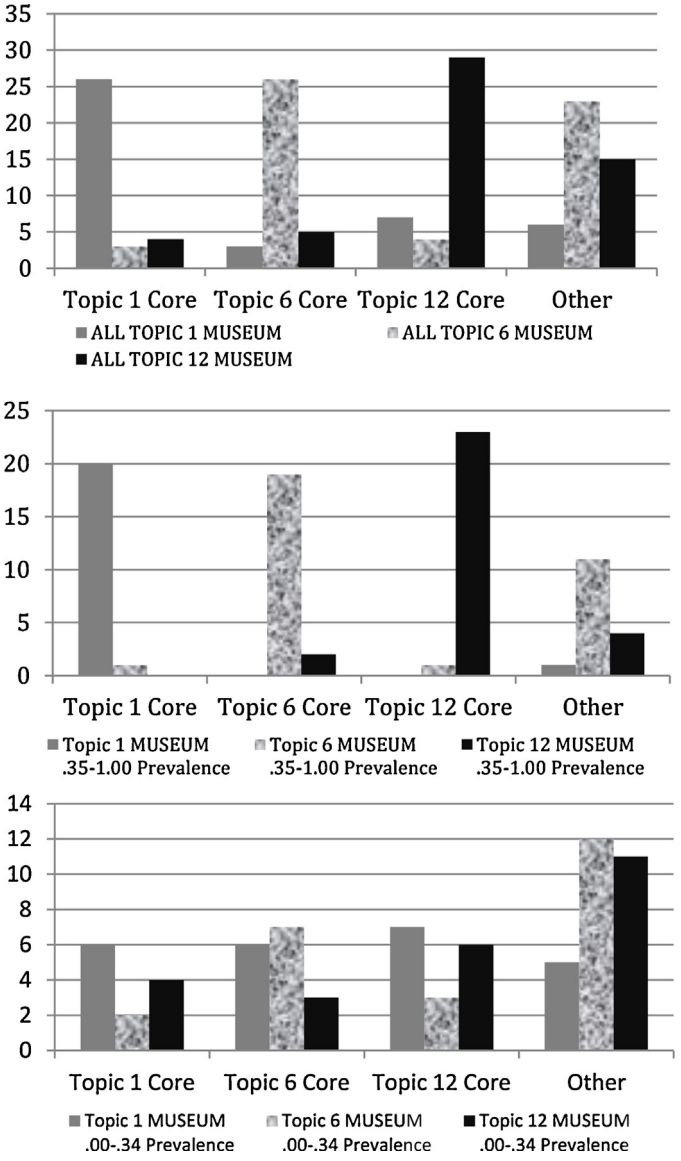

Fig. 2. Number of texts in which ''museum'' is assigned to Topic 1, Topic 6 and Topic 12, with meanings (core) consistent with the interpretation of those Topics for (a) all texts; (b) those with topic prevalence >.34–.89 (none was higher); and (c) those with topic prevalence <.35.

museum's relationship to its public. Finally, we interpreted Topic 12 as focusing upon exhibitions, so we treat its core as references to museums in the context of reviews of or stories about exhibitions. We classified as ''other'' references to museums that fit into none of these categories, such as stories about permanent collections, event listings, references to several kinds of arts organizations, philosophical discussions of art funding, references to museums in the

context of profiles of artists or others who had worked at or been exhibited at museums, references to the Institute of Museum Services, a story about an employee-discrimination lawsuit, and metaphorical uses of ''museum.''

Fig. 2's top panel demonstrates that the model satisfies the minimal condition for validation, that ''museum'' be used in different senses when assigned to Topics 1, 6 and 12, respectively. The model is validated by the very different senses in which ''museum'' is used when it is assigned to the three different topics (with a chi-square probability for the contingency table that is vanishingly small). When ''museum'' is assigned to Topic 1, 62 percent of the references are about the built environment, compared to fewer than 8 percent for Topics 6 and 12. When it is assigned to Topic 6, 46 percent of references are to grants or outreach, compared to fewer than 10 percent for the other topics. Finally, 55 percent of assignments of ''museum'' to Topic 12 concern exhibitions, compared to 17 percent for Topic 1 and 7 percent for Topic 6. These results largely validate our interpretations of the model.

When we distinguish between assignments to topics that account for many of the words in a text, on the one hand, and those that account for relatively few, on the other hand, we discover what may be an important limitation: discrimination among the senses of a term is far more successful when the topic is prevalent in a text than when that topic accounts for a small percentage of word assignments. The model distinguished extremely effectively among senses of ''museum'' for topics that were prevalent in texts. Putting aside the ''other'' category, the second panel of Fig. 2 indicates that in texts in which at least 35 percent of words were assigned to the relevant topic, the model was virtually unerring, assigning 20 of 21 references to the built environment to Topic 1, 19 of 21 references to grants or outreach to Topic 6, and 23 of 24 discussions of particular exhibitions to Topic 12. By contrast, as the third panel indicates, in texts with less than 35 percent prevalence, the model failed to discriminate significantly among different senses of ''museum.'' The problem may lie in the random assignment of particular instances of a word to topics, based on the topics' prevalence in that text. If a topic is represented sparsely, such random assignment may introduce a lot of noise. By contrast, if a topic dominates the text, words associated with it are likely to appear in suitable contexts. There is no magic to the crude dichotomization at 35 percent, of course, and the matter deserves further study.[12]

Although the results largely validated our interpretations, they did not do so completely. The value of testing for internal validity in this way is underscored because, in examining the disparity, we learned something important that leads us to amend our interpretation of Topic 6. Note the high representation of ''other'' codes in Topic 6, accounting for 34 percent of occurrences even in texts in which Topic 6 is relatively prevalent (considerably more than the 14 percent for Topic 12 and 5 percent for Topic 1). Seven of eleven references in the ''other'' category mention museums in the context of broad discussions of political philosophy or arts policy, often having to do with funding or with the role of the arts in society. By contrast, such contexts never occur for Topic 1 or Topic 12. In retrospect, the connection between these references and the emphasis in Topic 6 on grants and outreach—in effect instantiations of these broader issues—is understandable and leads to a broadened interpretation of Topic 6 as including general as well as specific dimensions of grant-making and community responsibility.

[12] The issue is potentially nontrivial. Fifty-one percent of word assignments are to texts in which a topic's prevalence does not exceed 35 percent, with a per-topic range of 37 percent (Topic 2) to 62 percent (Topic 9). Research should investigate the functional form of the relationship between prevalence and discriminatory power; thresholds at which discrimination becomes less effective; consistency of functional form and thresholds across models and corpora; and possibilities for improving discrimination by fine-tuning parameters in implementing the algorithm.

To summarize, we examined two terms (''museum'' here and ''film'' in Appendix B of the online supplement) that appeared prominently in more than one topic and compared the sense in which they appeared when assigned to each topic. We found that the models distinguished effectively between competing meanings of the terms. Although the discriminations were not perfect, almost all of the noise came from texts in which relatively few words were assigned to the topic in question. For texts in which a topic accounted for 35 percent or more of word assignments, topic models put into action the insight that meaning is relational, effectively discriminating between different senses of the same term in the absence of any semantic information. Moreover, the procedure validated our substantive interpretations of the topics, while providing information that led us to amend one such interpretation based on new information the analyses revealed.

5.2. *Heteroglossia*

Another central concept in the study of culture is *heteroglossia*, the capacity of a text to contain multiple voices and thus to speak in different ways to different audiences (Bakhtin, 1982 [1934–1941]). The notion of ''voices'' refers to characteristic modes of verbal expression (word choice, syntax, phrasing, and so on) associated with particular speech communities. Heteroglossia is related to polysemy in that voices may differ, in part, in their use of particular terms, thus introducing multivocality (the ability to speak in different ways to multiple audiences) into a text. A common example of such multivocality is ''dog-whistle politics''—the practice of inserting terms into political speeches that are intensely meaningful to subsets of listeners, but present more mundane meanings to the uninitiated.

Sometimes topics refer to particular subjects, but at other times they may refer to more than that: distinct voices in the text. Whereas *polysemy* refers to variations in meaning of a single term, *heteroglossia* refers to ambiguity at the level of the text. Identifying heteroglossia in texts is important for sociologists of culture, because doing so makes it possible to trace influence over time, as when phrases characteristic of one speech community (for example ''culture war'' language originating among religious conservatives) enter into texts produced by another (earnest liberals concerned about political polarization). Topic models are well suited to identifying heteroglossia because topic models are *mixed membership models*: Rather than assign texts to particular topics, they view texts as *mixtures of topics*. That is, a model is a mixture of topics that are shared across a collection of texts, with each document exhibiting the topics in different proportion. This is critical: Insofar as topics capture the voices of different speech communities, the distribution and co-occurrence of topics within texts represents a readily exploitable measure of heteroglossia.

The extent to which topics can be identified with particular speech communities, or ''voices'' in Bakhtin's (1982 [1934–1941]) sense, varies from corpus to corpus. It seems likely to be the case for literature (Bakhtin's own case) and for texts associated with particular scientific disciplines, political ideologies, or religious groups. By contrast, in news, writing stylistic idiosyncrasy is discouraged and journalistic norms of balance and neutrality suppress ideological variation.

Even in news writing, however, one can discern different voices in material that a reporter quotes or paraphrases. We can do no more than give a brief example here, an example that will demonstrate clearly the multiple-membership feature of topic models and will at least suggest (without providing sufficient background or analysis to confirm our interpretation) the way in which this feature may contribute to the analysis of heteroglossia.

Fig. 3 illustrates how texts are parsed among different topics. It depicts an article reporting on the announcement of several NEA grants. The box on the upper left corner lists topics to which

words in the article have been assigned. In the standard output, colors are used to signify words assigned to each topic. Here we use two shades of gray to illustrate terms assigned to Topic 2 (controversial grants [light shading]) and Topic 6 (education, outreach and multiple grantees [dark shading]), which together account for 90 percent of assignments.

Consider two sentences from the article in Fig. 3:

"The Texas Commission on the Arts received a $50,000 grant for developing arts programs in rural, underserved or inner-city areas of the state. That grant is part of the NEA's new push to encourage the arts in areas that historically have had few artistic opportunities."

"It seems to me that the only reason that I, as chairman, could throw it out and substitute my own judgment would be if I were really persuaded that there was no evidence of artistic substance in the proposed grant,' he said."

Although they appear in the same article, these two sentences are dominated by different topics. More than 70 percent of assigned words in the first passage are assigned to Topic 6 and 80 percent of assigned words in the second are assigned to Topic 2. These, we would suggest, represent different voices, or perhaps different timbres of a bureaucratic voice.

The first includes terms that arts agencies imported from the public social-service delivery system. (In the late 1970s, some arts policy makers went so far as to speak of the "arts service delivery system" [DiMaggio, 1986, p. 6].) The service-provision voice is bold ("NEA's new *push*"); it is about "*developing programs*;" it defines the public as a set of constituencies organized around community type ("*rural*. . .or *inner-city areas*"); and it emphasizes values of equity and justice ("*underserved*. . .*areas,*" "*artistic opportunities*").

By contrast, the second excerpt is written in the register of bureaucratic justification, making reference to *reason, persuasion*, and *evidence*, as well as *artistic substance* (a careful reference to the obscenity standard of *Miller v. California*, to which defenders of the Endowment were gravitating [Fiss, 1991]). An earlier passage in the article, also consisting primarily of words assigned to Topic 2, emphasizes process, referring to the panels of "citizens," in contrast to the first passage, viewed as a universal role rather than partitioned into communities, to whom proposal review is delegated.

This example demonstrates several things. First, it illustrates the fact that texts are mixtures of different topics. In this case, Topic 2 was dominant and Topic 6 subdominant. Second, it shows that passages that are dominated by different topics may embody different voices, different modes of expression defined by word choice (reference to programs and constituencies or reference to choices and responsibilities), emotional tone (bold or defensive), forms of justification (distributive equity vs. conformity to procedure) and values (opportunity or artistic substance), as well as, in certain cases, elements of syntax and grammar. Note that passages receive their coloration from the topic that dominates them: not every relevant word must be assigned to that topic.[13]

Third, while topics may embody different voices, they are not necessarily coterminous with them. Even within this article, Topic 2 features at least two voices—the style of bureaucratic

[13] Note that the model assigns topic distributions to each text, but within texts assigns words to topics randomly, based on these distributions, without reference to their location within the text. Designing models in which assignments are sensitive to intra-textual context (i.e., that transcend the "bag-of-words" assumption) would help to fine-tune the analysis of heteroglossia, especially for topics that are weakly represented.

Text for Article 349

Topic	Proportion
1	0.00359712230215827
2	0.672661870503597
3	0.0359712230215827
5	0.0539568345323741
6	0.230215827338129
7	0.00359712230215827

Section: HOUSTON
Headline: Arts endowment announces grants
Pub Date: 11/06/1991
Pub Name: Houston Chronicle
Page: 6

Corpus Text:

Arts endowment announces grants

WASHINGTON -- The National Endowment for the Arts announced Tuesday 735 grants to artists and organizations, totaling $16.85 million for the final quarter of the current fiscal year.

Included are $8,000 fellowships to two performance artists who are suing the endowment because their performances about gay life were turned down before.

Texas cities collected 17 grants, totaling $227,230.

The Texas Commission on the Arts received a $50,000 grant for developing arts programs in rural, underserved or inner-city areas of the state. That grant is part of the NEA's new push to encourage the arts in areas that historically have had few artistic opportunities.

The Texas total includes three Houston grantees: solo recitalist Isabelle Ganz, a mezzo-soprano, $12,000; museum professional Jeannette Dixon, $7,000; and the arts organization Concerned Musicians of Houston, $10,000.

Controversial performance artist Tim Miller, of Santa Monica, Calif., said that it ""remains to be seen" whether he and fellow artist Holly Hughes will pursue their lawsuit against the NEA now that they have received the new grants.

""It is certainly cheering that the grants to Holly Hughes and me went through, but I feel the chilling effect is still very much there to artists," Miller said.

Miller said he would use the grant to develop performance pieces dealing with ""the issues of my life as a gay person and the AIDS crisis."

Hughes, who lives in New York, could not be reached.

Hughes and Miller, as well as the other members of the so-called ""NEA Four," have performed at Houston's DiverseWorks in the past two years.

John E. Frohnmayer, chairman of the endowment, said Miller and Hughes were among 19 performance artists who received grants, out of 124 applicants. They were approved by a panel of private citizens, most of them artists, by the presidentially appointed National Council on the Arts and by the chairman.

Frohnmayer said Miller ranked second and Hughes tied for fourth in a ranking of the applicants by the panel.

""It seems to me that the only reason that I, as chairman, could throw it out and substitute my own judgment would be if I were really persuaded that there was no evidence of artistic substance in the proposed grant," he said. ""And in these two along with every other grant we are announcing, I am persuaded there is artistic value.

""I would hope these individuals would not be condemned for who they are," he said.

The NEA chairman said he alerted the White House that the announcement was coming. ""Nobody in the Bush administration suggested that I should or should not approve the grants," he said.

Frohnmayer vetoed performance artist grants for Miller, Hughes, Karen Finley and John Fleck in June 1990. The endowment said his action was taken on artistic grounds. The four have challenged the action in U.S. District Court in Los Angeles, arguing they were rejected for political reasons.

According to a transcript of a closed meeting of the national council, Frohnmayer cited performances by Finley in which she inserted vegetables into bodily orifices and by Fleck in which he urinated on stage.

Both the House and the Senate adopted an amendment to an appropriation bill this year to prohibit the government from financing ""patently offensive" sexual exhibits, but a conference committee eliminated the language from the bill, which awaits President Bush's signature.

This quarter's NEA grantees include 212 individual artists, 60 folk arts organizations, 172 other arts organizations and 38 state arts agencies.

Fig. 3. Sample article demonstrating assignment of words to topics.

justification noted above and references, noted elsewhere in the article, to rights language and narratives of censorship and political struggle (in passages quoting aggrieved artists).

Our purpose here is to illustrate the *potential* of topic modeling for advancing empirical research on heteroglossia, not to convince the reader of our interpretation. A full analysis would make a stronger substantive case for the existence of particular voices. It might well identify key elements of these voices by analysis of "training texts"—exemplary texts known to embody particular modes of expression. The results of these analyses, which would employ supervised approaches to topic modeling (Blei and McAuliffe, 2007), could then be used to identify such voices in a corpus of unknown texts.

5.3. Topics as frames and predictive validation of the model

A central concept in the sociology of culture is the interpretive "frame" (Snow, 2004). A frame is a set of discursive cues (words, images, narrative) that suggests a particular interpretation of a person, event, organization, practice, condition, or situation. According to Gamson et al. (1992, p. 384), "frame plays the same role in analyzing media discourse that schema does in cognitive psychology—a central organizing principle that holds together and gives coherence and meaning to a diverse array of symbols." From the perspective of social cognition, frames employ condensed images in ways that prime particular schemas, and the networks of associations they entail (Johnston, 1995). For example, texts that refer to government grants to local museums or government support for arts classes in local schools are likely to evoke different (and more positive) associations than texts that refer to charges that a public agency has supported blasphemous or pornographic artwork. Issue framing may be intentional (as is ordinarily the case in politics or advertising) or it may occur without strategic intent (as is often the case in journalism).

With some exceptions, empirical work on framing has lagged behind theoretical development (Benford, 1997). Topic modeling provides a promising approach because the sets of terms that constitute topics index discursive environments, or frames, that define patterns of association between a focal issue and other constructs. When applied to corpora that cover particular issue domains (like government funding for the arts), topic modeling has some decisive advantages for rendering operational the idea of "frame" in media research—such as facilitating analysis of larger corpora than human coders can master, facilitating discovery of unanticipated frames, and distinguishing between different uses of the same term. As a multiple-membership model, topic modeling is especially useful for discerning frames in press accounts, which typically incorporate multiple frames (Benson, 2013, p. 4). After topics have been identified, *frame-specific* counts of the prevalence of particular terms or analysis of the relationship between frames and other text features may be a useful elements in the interpretive process.

Here is an example from the arts-funding model. Recall that we are primarily interested in three topics that highlight political and social conflict over government arts support. Each can be viewed as a frame, in that it includes terms that call attention to particular ways in which such support may arouse controversy: a controversial-art frame focuses on debates over particular grants; a legislative-conflict frame refers to congressional debates over the NEA; and a culture-war frame treats arts funding as one of many related issues in a broader cultural struggle. Each of these conflict frames represents a negative discursive environment for government arts programs as compared to frames that emphasize the contribution of such programs to events and institutions of which the public generally approves.

Fig. 4 examines change over time in the discursive environment by aggregating the percentage of all words assigned to the three conflict frames as compared to the percentage assigned to three

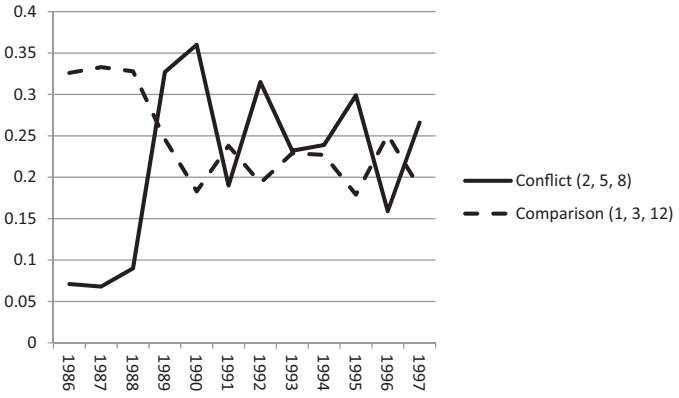

Fig. 4. Percentage of words assigned to conflict frames vs. comparison frames, 1986–1997.

comparison frames (those associating arts funding with urban development, musical events, or [uncontroversial] art-museum exhibits). From 1986 through 1988, the comparison topics vastly overshadow the conflict topics, accounting for a stable 33 percent of word assignments in each year, compared to between 7 and 9 percent for the three conflict topics. In 1989, we see a dramatic and sudden shift in framing, with 33 percent of all words assigned to conflict terms in 1989 and 35 percent in 1990. Although the percentages fluctuated thereafter, the conflict topics accounted for more than 15 percent of the total through 1997 and for more than 20 percent in all but two years. Thus, Fig. 4 demonstrates a marked deterioration in the discursive environment for arts funding after 1988.

We can also use the timing of increases and declines in the prevalence of particular topics as a means of evaluating external validity. If we interpret the conflict topics correctly as frames associated with controversial arts grants, congressional turmoil, and political and cultural polarization, these topics should become more prevalent in response to certain events. For example, if our interpretation of Topic 2 as being about arts controversies is accurate, its prevalence should increase when controversial grants are in the news. If our interpretation of Topic 5 as being about congressional conflicts over the arts is accurate, we would expect it to ebb and flow with the congressional budget cycle and to be particularly prominent when Republicans controlled one or both chambers. If our interpretation of Topic 8 as embedding the arts in a broader discourse of cultural cleavage is correct, more words should have been assigned to Topic 8 during Patrick Buchanan's primary campaign against Pres. George H.W. Bush and during periods when press attention to the so-called "culture war" was high.

Consider the following hypotheses (numbered according to the topics they concern), each of which predicts variation in the monthly number of words assigned to a topic. For each hypothesis, Table 3 reports the effect of hypothesized predictors, controlling for the number of words assigned to all topics *except* the conflict topics. The latter adjusts for the amount of attention given the arts overall (aside from contention over public funding) and the size of the newspapers' news holes (space available for news content). For each hypothesis we report the R squared statistic and the significance of the hypothesized effect. For example, the first row indicates that Hypothesis 2a is being tested and that significantly ($p < .001$) more words (5919 per month) were assigned to Topic 2 in months in which we expected Topic 2 to be more prevalent. The overall amount of attention to uncontroversial arts topics had a positive but insignificant effect,

Table 3
External-validity hypothesis tests (OLS regression with words assigned to topic monthly as dependent variable).

Hypothesis	Period dummies	Other topics	Cyclical dummies	Conflict phrases	Adjusted R^2
2a: 1/86-3/89 v. 4/89-11/89	5919**	.123			.631
2b: 4/89-10/90 vs. *11/90-12/97*	5051**	.168**			.613
5a: 4/89-11/94 vs. *12/94-12/97*	809*	.151**			.156
5b: Annual budget cycle		.149**	924*		.192
5a & 5b together	860*	.145**	945*		.223
8a: texts/month with conflict phrases		.170**		578**	.273
8b: 8a+ dummies for 1992 and fall 1994	2193**	.183**		539**	.397

* $p \leq .05$.
** $p \leq .001$.

with both predictors explaining just under two thirds of the variance in the number of words assigned to Topic 2 in each month.

Hypotheses were based on histories of the NEA and policy analyses (Alexander, 2000; Burgess, 2006; Dubin, 1992; Fiss, 1991; Frohnmayer, 1993; Jensen, 1995; Kimbis, 1997; Koch, 1998; Kresse, 1991; Ziegler, 1994) and not on the newspaper articles in our corpus. By predicting when the prevalence of particular topics should increase based on information external to our study, we can test the external validity of our interpretations.

Hyp. 2a. *Topic 2 should be more prevalent in April through November 1989 than before April 1989*. There were no significant arts controversies around NEA funded works between 1986 and early 1989. The ''Piss Christ'' controversy erupted in April 1989 and conservative legislators joined the protest in May. In June 1989, the Corcoran canceled the Mapplethorpe exhibit and criticism of the Endowment continued. Chairman John Frohnmayer's requirement of a decency pledge for grantees and cancelation of several controversial grants extended the controversy into the fall. This hypothesis is confirmed (see first row of Table 3), with the model explaining 63 percent of variation in monthly Topic 2 word assignments over this period.

Hyp. 2b: *Topic 2 should be less prevalent after October 1990 than between April 1989 and October 1990*. Controversy continued (with only a brief let-up in late fall 1989) through October 1990, with an art-museum director arrested for presenting the Mapplethorpe exhibit in Cincinnati in April, attacks by conservatives on NEA grants to gay-themed productions and exhibits later that spring, more vetoes of controversial grants in June, debates over decency requirements and a lawsuit against NEA by four performance artists. After October 1990, much of the action moved to Congress, which authorized a Commission to study the Endowment and devised a series of legislative initiatives to control the agency or reduce its funding. This hypothesis is also confirmed (row 2 of Table 3), with 61 percent of the variance in monthly Topic-2 word assignments for the relevant period explained.

Hyp. 5a. Topic 5 should be more prevalent after the Republican takeover of Congress in November 1994 than between April 1989 and November 1994, when the Democrats controlled Congress. We interpret Topic 5 as focusing upon Congressional debates over arts funding, and interpret the rise in Topic 5's prevalence as reflecting a shift in the action from the press and conservative movement groups to Congressional hearings and formal legislation. If this is the case, then we would expect Topic 5 to be more prevalent when Republicans controlled Congress, as this enabled them to influence the legislative agenda and committee process. Because Republicans attained majorities in Congress during the off-year elections of November 1994, we predict that,

other things equal, more words will be assigned to Topic 5 after November 1994 than earlier in the controversy (Spring 1989 through November 1994). The hypothesis is supported (see row 3 of Table 3), with 809 more words assigned to Topic 5 in the typical month after November 1994.

Hyp. 5b. After 1988, Topic 5 should be more prevalent between April and October (except for August) than between November and March. In addition to the historical shift marked by Hypothesis 5a, we anticipate that attention to Topic 5 will have a cyclical component. Many of Congress's efforts to discipline the NEA (which did not get under way until 1989) were attached to authorization or, more often, appropriations bills, which tend to be active between April and October (with the exception of the recess month of August). The timing is imprecise: when Republicans controlled Congress, budget politics started as early as February or March and, during some years, extended through November. Nonetheless, this hypothesis is supported (line 4 of Table 3): On average, 924 more words were assigned to Topic 5 in April, May, June, July, September and October than in other months. When the cyclical measure is included in the same model as the historical measure (see line 5), the effects of both increase, as does variance explained.

Hyp. 8a. Topic 8 should be more prevalent in months in which the press uses phrases like ''culture war'' that index wide-ranging cultural contention. We have interpreted Topic 8 as embedding controversies over arts funding in a broader rhetoric of cultural contention, polarization, and moral decline. Whereas Topic 2 identifies controversial grants as problems in themselves, Topic 8 views them as instances of broader cultural trends, associated with such other issues as abortion, sexual media content, and homosexuality. If this interpretation is correct, then the number of words assigned to Topic 8 in any month should be a function of the incidence of terms that index broader themes of cultural contention. As indicators, we used ''culture war'' (which appeared in fifty-six articles), and ''moral decline'' and ''moral decay'' (each of which appeared in two). These compound terms were not identified as entities in the modeling stage and therefore did not influence the topic model results. The hypothesis receives strong support: for every text per month in which one of these phrases appears, the number of words assigned to Topic 8 increases by 578 ($p < .001$).

Hyp. 8b. The prevalence of Topic 8 should increase during electoral campaigns in which prominent candidates built campaigns around social conservative issues. Specifically, this includes the 1992 Republican campaign in which columnist Patrick Buchanan challenged George H.W. Bush, declaring a ''culture war'' on secular humanists, and the 1994 off-year elections in which Republicans regained control of Congress with strong support from the religious right. Thus we predict that, controlling for the prevalence of non-conflict topics, dummy variables for 1992 and for the fall 1994 election season will independently increase assignment of words to Topic 8. This hypothesis is supported: when these month dummies are added to a model including the cultural-conflict terms, they increase the monthly assignments to Topic 8 by 2194, slightly reducing the impact of each text with a conflict phrase and explaining 40 percent of variance.

These analyses support our interpretations of Topic 2, Topic 5, and Topic 8. Although we are not surprised, we believe that such external-validity checks represent a useful step in topic-model analysis. Remember that the program that produced the topics used no information about the meaning of the words in the texts or about the political context in which the texts were produced. The topics were only endowed with meaning post hoc when we interpreted them. By testing such interpretations against expectations generated from information external to the study data, researchers defend against the temptation to over-interpret LDA term lists and then cherry-pick examples of texts that support their interpretations. Once one has validated one's interpretations in this way, data on topic prevalence can be used reflexively as evidence about the state of the world.

6. Substantive application: did news outlets vary in their use of conflict frames?

For students of culture, producing a topic model is not an end in itself, but rather the first step in a longer process of interpretation and analysis. In the current case, by construing Topic 2 (arts controversies) and Topic 8 (polarization) as alternative frames for depicting conflict over arts grants—in one case treating controversial grants as a distinct problem, in the other treating arts funding as one of many related ''social issues''—we can ask whether different news outlets varied in their framing of this issue.

Several possibilities suggest themselves. Did our conservative paper, *The Wall Street Journal*, emphasize conflict frames to a greater degree than the other outlets?[14] Did the relatively liberal *Times* and *Post* pay less attention to the attacks on the NEA? Or did they feature them more prominently because the *Times* is the hometown paper of much of the art world and the *Post* views congressional coverage as a major part of its mission? Did the regional papers' patterns of coverage reflect their local art scenes (free-wheeling performing arts in Seattle, major art collections in Houston) or social environments (more liberal in Seattle, more socially conservative in Houston)? We approach these questions in several ways.

6.1. The mutual information criterion: gaining information about differences among sources from topic validation

We use a new Bayesian model-checking technique (Mimno and Blei, 2011) to ask to what extent the placement of words in texts is consistent with the model assumption that words assigned to a topic are drawn independently from the same multinomial distribution, and to learn from deviations from this assumption. In this example, we focus on differences among sources in the co-occurrence of particular terms within topics (with a complementary analysis of differences associated with particular texts reported in Appendix C of the online supplement). Our strategy here is to examine how the topic model *misfits* the data. We emphasize that ''misfit'' is not a bad thing—all models misfit the data. Part of the process of exploratory analysis is to dig into, and learn from, where the misfits occur.

We analyze, per-word, when the independence assumptions are violated by the model. In particular, we use ''mutual information'' (MI), an information-theoretic measure of how related two random variables are. In theory, LDA assumes that the observed words assigned to a topic are drawn independently from that topic's distribution. Those observations should thus be independent of an external variable. Examining how and when this is true—by measuring the independence to a variety of external variables—gives us more insight into the texts (Mimno and Blei, 2011).

We focus here on deviations from independence based on newspaper source. We simulate each term's mutual information under true independence and compare it to the observed mutual information. This approach is an example of posterior predictive checking (Gelman et al., 1996; Rubin, 1984), a more general methodology for examining the fitness of Bayesian models. Each panel in Fig. 5 presents the observed MI scores and expected values for each topic (one topic per

[14] Based on editorial policy on Supreme Court decisions between 1994 and 2004, Ho and Quinn (2008) classify the *Times* as very liberal, the *Post* as somewhat liberal, the *Houston Chronicle* as centrist, and the *Wall Street Journal* as conservative. (These judgments were not absolute, but were based on comparison among twenty-five newspapers.) Based on the frequency in news content of phrases articulated by Republican and Democratic legislators, respectively, Gentzko and Shapiro (2007), using data from 2005, classify the *Times* and *Post* as relatively (and comparably) liberal and the *Seattle Times, Houston Chronicle*, and *Wall Street Journal* as relatively (and similarly) conservative.

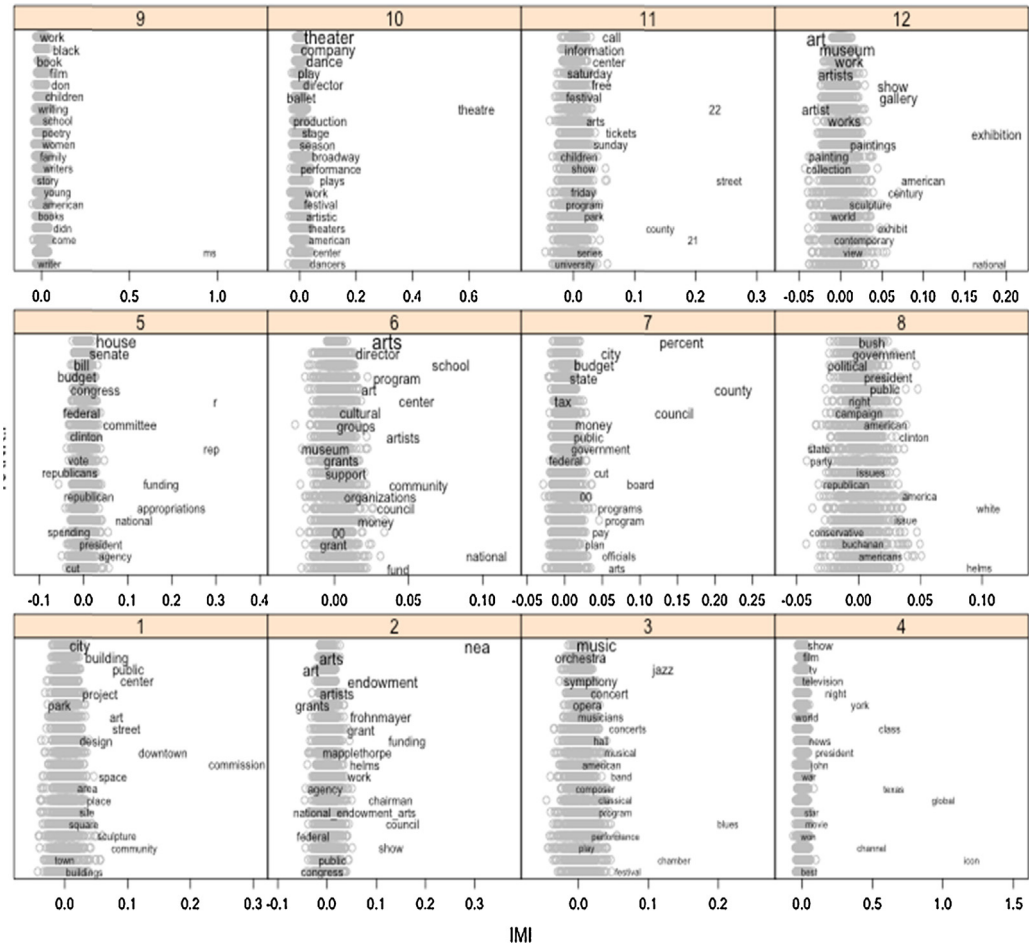

Fig. 5. Instantaneous mutual information: top 20 terms from twelve topics grouped by source.

cell) from a procedure entailing 100 resamples from the posterior. In each of the two panels, each cell refers to a different topic (the number of which appears at the top of the cell). Each horizontal line within each cell represents observed MI values (marked by the first letter of the word) for the top 20 terms associated with that topic. Predicted values from the resamplings are represented as gray circles. The position of each term relative to the gray line indicates the degree of association with other terms (mutual information). The further to the right the word appears, the greater the amount of mutual information and the larger the difference between the observed value and that predicted under the assumption of independence. The results thus indicate whether certain terms within topics appear in particular newspapers more than one would expect by chance.[15]

Some differences are easily explained. In Topic 7, for example, the appearance of ''percent'' well to the right of the gray line reflects the fact that only some cities have ''percent-for-art'' laws requiring art installations in major building projects. The position of ''commission'' in Topic 1 no

[15] Such analyses can point in the direction of more complex topic models, such as hierarchical variants by source, dynamic topic models that account for time, or bursty topic models (Doyle and Elkin, 2009).

doubt reflect the important role in Seattle of the Kings County Arts Commission. The appearance of ''ms'' well to the right in Topic 9 may reflect the *Times* policy of preceding names with titles.

Other differences hold more substantive interest. For example, the location of terms in the Topic 3 cell suggests that when newspapers covered music they were similarly interested in symphonic music and opera, but varied in attention paid to blues, chamber music and jazz. Another example: Note the location of ''NEA'' and ''endowment'' outside their corresponding-gray bands under Topic 2. This suggests that in writing about controversial art works, some sources may have focused more on the NEA's role than did others. Similarly, the location of several Topic 5 terms indicates that some sources placed more emphasis on actions of congressional committees than others in discussions of congressional arts politics.

We conducted additional analyses (results available on request) inspecting changes in mutual information scores when, instead of categorizing all texts by source, we broke down sources into two categories, one for one newspaper and one for all others. We did this *seriatim* for all five sources: MI scores for a topic would remain high or increase where a particular source differed markedly from the others. They would decline when a source was similar to others in its distribution of terms within that topic. These additional tests indicated that The *New York Times* was distinctive in its attention to the NEA, perhaps reflecting the fact that many of the artists and institutions involved in legal or public relations battles with the NEA were located in New York; and the *Washington Post* accounted for much of the covariation of terms and sources for Topic 5, no doubt due to its especially intensive coverage of Congress.

Taken as a whole, these analyses tell us, first, that, as expected in any corpus of this kind, results diverge from the independence assumptions of the model, and that some of this divergence is related to differences among newspapers. In the next section, we pursue this insight further and ask how particular newspapers differed not just in their representation of terms within topics, but in the representation of the topics themselves within the texts. In particular, we look for substantive differences in the incidence of word assignments to Topic 2 (arts controversies), Topic 5 (congressional art-policy actions), and Topic 8 (polarization) relative to other topics.

6.2. *Variation among sources in prevalence of three conflict frames*

In the analyses that follow, the unit of analysis is the text, the percentages of words in each text assigned to each topic are the dependent variables, and the independent variables are the source of the text (i.e., a set of news source dummies [*NY Times* omitted]), year of publication [also a series of dummies], and the text's length in words). We focus upon coefficients indicating the impact of each source (net variation associated with year or with the tendency for some topics to be featured in longer articles) on the relative prevalence of the arts-controversy frame (Topic 2), the legislative-action frame (Topic 5) or the culture-wars frame (Topic 8), as compared to other topics. To do this we employ the fractional multinomial logit model (FMNL) in STATA (Buis, 2008), a model similar to the multinomial logit but designed to predict probabilities (or percentages) summing to 1.

Table 4 summarizes results, comparing each newspaper to the *New York Times* (the reference category). The left panel contains comparisons for Topic 2 (the arts-controversy frame). The center panel includes comparisons for Topic 5 (congressional-action frame). The right panel includes comparisons for Topic 8 (the culture-wars frame). The entries in the left column of each panel indicate, for each row, which topics are being compared. Entries in the first row indicate the extent to which each newspaper (column headings) assigned more or fewer words to Topic 1 relative to the focal topic than did the *New York Times*. Heavily shaded cells indicate that a

Table 4
Fractional multinomial logit analysis predicting relative prevalence of Topic 2 (left panel), Topic 5 (center panel), and Topic 8 (right panel) compared to reference topic (indicated on row), each newspaper compared to *New York Times*; coefficients represent impact of source, controlling for year and word count.

Topic 2

	HOUSTON	WAPO	JOURNAL	SEATTLE
1 vs. 2	0.0671	-0.156	-0.142	1.32
3 vs. 2	-0.027	-0.383	-0.773	0.462
4 vs. 2	0.612	0.159	0.56	0.661
5 vs. 2	0.101	0.231	0.767	0.207
6 vs. 2	0.377	-0.123	-0.984	0.611
7 vs. 2	0.369	-0.031	0.561	0.845
8 vs. 2	0.218	-0.13	0.932	0.209
9 vs. 2	-0.134	-0.015	0.025	0.613
10 vs. 2	-0.16	-0.589	-0.939	0.311
11 vs. 2	0.02	-0.054	-1.767	0.709
12 vs. 2	-0.569	-0.659	-0.561	-0.05

Topic 5

	HOUSTON	WAPO	JOURNAL	SEATTLE
1 vs. 5	-0.034	-0.388	-0.909	1.116
2 vs. 5	-0.101	-0.231	-0.767	-0.207
3 vs. 5	-0.128	-0.614	-1.54	0.255
4 vs. 5	0.511	-0.072	-0.207	0.454
6 vs. 5	0.275	-0.354	-1.75	0.404
7 vs. 5	0.268	-0.262	-0.206	0.638
8 vs. 5	0.116	-0.362	0.165	0.002
9 vs. 5	-0.236	-0.246	-0.743	0.407
10 vs. 5	-0.261	-0.82	-1.706	0.104
11 vs. 5	-0.081	-0.287	-2.534	0.502
12 vs. 5	-0.67	-0.89	-1.328	-0.257

Topic 8

	HOUSTON	WAPO	JOURNAL	SEATTLE
1 vs. 8	-0.151	-0.026	-1.074	1.114
2 vs. 8	-0.218	0.130	-0.932	-0.209
3 vs. 8	-0.190	-0.252	-1.705	0.253
4 vs. 8	0.395	0.290	-0.371	0.452
5 vs. 8	-0.116	0.362	-0.165	-0.002
6 vs. 8	0.159	0.008	-1.916	0.402
7 vs. 8	0.151	0.100	-0.371	0.636
9 vs. 8	-0.352	0.116	-0.907	0.404
10 vs. 8	-0.377	-0.458	-1.871	0.102
11 vs. 8	-0.198	0.075	-2.699	0.500
12 vs. 8	-0.787	-0.529	-1.493	-0.260

Note: Dark gray shading = Significantly more attention to focal topic (i.e., the topic to which others are compared) than NY Times. No shading = Significantly less attention to focal topic than NY Times. Moderate shading = No difference from NY Times. Cell entry = multinomial logit coefficient. (Positive coefficient means less attention to focal topic relative comparison topic than NY Times; negative coefficient means more attention to focal topic. Models include controls for year and word count.

newspaper employed the focal frame (relative to the comparison topic) significantly *more* than did the *New York Times.* Unshaded cells indicate that the newspaper employed the focal frame significantly *less than* did the *New York Times*. Moderately shaded cells indicated no significant difference. For example, the upper-right-hand cell of the first panel indicates that articles published in the *Seattle Times* assigned a significantly smaller percentage of words than the *New York Times* to Topic 2, as compared to Topic 1 (controlling for article length and publication year).

The left panel of Table 4 shows the *Seattle Times* to be an outlier, assigning significantly fewer words to Topic 2 (arts-controversy frame) than the *New York Times*, relative to ten of eleven other topics. (The exception, Topic 10 [theater and dance] reflects the *New York Times* intense coverage of two art forms in which New York is pre-eminent.) The center and right panels indicate that the *Seattle Times* also allocated significantly fewer words to Topic 5 (legislative actions) and Topic 8 (polarization frame) than did the *New York Times* relative to over half of the other topics. Thus, the *Seattle Times* referred more frequently to arts funding in connection with uncontroversial artistic performances and exhibits and with arts' contribution to urban development, and less often with reference to political controversy or social conflict than did other newspapers. This pattern may have reflected an especially cordial relationship to Seattle's vibrant arts community or a socially liberal readership or both.

The *Wall Street Journal*, the most editorially conservative newspaper of the group, was an outlier in the other direction, assigning more words to Topic 8 (polarization) than to ten of the eleven other topics compared to the *New York Times* or most of the other papers. (The only exception was another conflict topic—Topic 5 [congressional politics]—on which the *Journal's* emphasis [relative Topic 8] was similar to that of the *Times*.) Indeed, the *Journal* emphasized all

conflict frames more than other papers. The *Journal* also focused significantly more on Topic 5 than did the *Times* compared to eight of the other eleven topics (the exceptions being Topic 8, of course, as well as Topics 4 [commercial media] and 7 [local and state budgetary issues], both topics that the *Journal* tended to assign with higher frequency than other sources). It also emphasized Topic 2 (arts controversies) more than did the *NY Times* in comparison to Topics 3, 6, 10, 11, and 12 (music, education and outreach, theater and dance, event information, and visual arts) but less than the *Times* compared to the two other conflict topics and Topics 4 (media) and 7 (local government finance). Consistent with these results, an analysis using the STATA fmlogit procedure (results available upon request) indicated that the *Journal* assigned between 5.2 and 6.4 percent more words to Topic 8 per text than other sources, after controlling for year and word count, making it an extreme outlier. The *Journal* was also an outlier on Topic 5, assigning between 3.0 and 5.1 percent more words to this topic than the other papers. (By contrast, *Journal* articles were assigned only slightly more topic 2 words.)

The *Journal's* heavy use of the cultural-wars frame, a theme popular among conservatives (DiMaggio, 2003), and its intense attention to congressional criticism of the NEA, is consistent with, although it does not prove, the possibility that the *Journal's* coverage of arts funding reflected its conservative politics. Also consistent with this interpretation, the *Journal* emphasized topic 2 (controversies) more during the Democratic Clinton administration than during the Republican administration of George H.W. Bush.[16]

The other two newspapers, the *Houston Chronicle* and the *Washington Post*, deviated only modestly and unsystematically from the *NY Times*. The Houston paper assigned more words to Topic 2 than the *NY Times* relative to Topic 12 (art exhibits), but fewer relative to Topics 4, 6, 7 and 8. Similarly, the *Chronicle* emphasized the polarization frame (Topic 8) more than the *Times* relative to three topics (9, 10, and 12) but less relative to three others (4, 6, and 7). The *Washington Post* assigned more words than the *Times* to both Topics 2 and 8 relative Topics 3 (music), 10 (narrative forms), and 12 (art exhibits), but fewer relative to Topic 5 (congressional action) and, for Topic 8 only, Topic 4 (mass media). The *Post* also assigned more words than the *Times* (though, for most comparison topics, less than the *Journal*) to Topic 5 (legislative action), a discrepancy that is unsurprising because Congress constitutes an important part of the *Post*'s hometown beat (Burgess, 2006, p. 113).

In sum, then, newspapers framed government arts funding differently, with the *Wall Street Journal* emphasizing conflict frames and the *Seattle Times* tending to eschew them, compared to the *New York Times, Houston Chronicle*, and *Washington Post.* Differences in topic prevalence were driven by both the stories the papers covered and the ways they covered them, due to varying missions (the *Journal*, a national business daily, referred more to commercial media and used fewer terms related to event listings), different news beats (the *Times*, in arts-rich New York, assigned more words to topics related to theater and art, whereas the *Washington Post* devoted focused more on Congress), and, possibly, differences in political orientation.

7. Conclusions and further work

This article describes how to use probabilistic topic models of newspaper articles to study cultural trends, moods, and depictions. We studied press coverage of government grants supporting the arts between 1986 and 1997. During this period, such grants became controversial and the National Endowment for the Arts, the federal agency that made many of them, faced

[16] Analyses available upon request.

fierce attack from Republicans and conservative social movement organizations. In this section we first summarize key substantive findings and then reflect on implications of the approach for other research on culture.

7.1. What did we learn?

Our analyses provide substantive insight into the response of the press to political attacks on the National Endowment for the Arts. Using LDA to discover themes and additional methods to exploit the results of the topic models, enabled us to gain the following insights:

(1) The tone of press coverage of arts funding shifted dramatically in 1989 from largely celebratory to substantially focused on controversy, producing a cloud of negative representations that persisted to varying degrees throughout the 1990s.
(2) Negative coverage of the NEA emerged suddenly with the election of George H.W. Bush (even though the controversial grants had been made during the administration of Ronald Reagan), consistent with the possibility that the attacks were part of an internal struggle between social conservatives and moderates for control of the Republican party.
(3) Press coverage reflected three different frames for the controversy: a focus on objectionable grants as bureaucratic errors; a focus on congressional efforts to punish the NEA for its mistakes; and a view of debates over the NEA as one of many fronts in a larger culture war. The second of these reflected both the budget cycle and the extent to which Republicans influenced the congressional agenda. By contrast, the third frame displaced the first over the course of the 1990s, as debates over the arts were integrated into a larger narrative about cultural polarization.
(4) Newspapers varied in their coverage of government arts patronage. In particular, the *Wall Street Journal*'s coverage focused on controversy more than did that of the other sources and, in particular, emphasized the culture-wars frame, especially after the election of Bill Clinton, while the *Seattle Times* emphasized more positive stories about government grants in support of local projects and institutions.

7.2. Using topic models to study culture

Throughout, we have emphasized the *fit between theory and method*, illustrating the way in which topic models render operational such concepts as frames, polysemy, heteroglossia and the relationality of meaning. Topic models provide an opportunity for sociologists of culture to add empirical substance to their analytic efforts by operationalizing key concepts using large textual corpora. At the same time, they present novel problems of interpretation and validation. To social scientists accustomed to analyzing population data, topic models (and other techniques to reduce the complexity of large corpora) may be perplexing. Rather than drawing inferences about a known population from a sample, the analyst is taking a population—all of the texts in the corpus—and must evaluate alternative accounts of its structure. The standard for selecting a solution is not so much accuracy as utility: Does the model simplify the data in a way that is interpretable, passes tests of internal and external validity, and is useful for further analysis?

This has several implications:

(1) The model is just the beginning. For cultural analysis, the purpose of modeling is to apprehend the structure of the data and render it tractable by producing meaningful topics (interpretable, depending on the data, as voices or frames) that can be used to answer more

focused questions. In this article, we used the topic model, first, to track change in representations of arts funding over time and, second, to ask whether different newspapers depicted arts funding in different ways. To answer the latter question, we analyzed data that the topic model produced with conventional regression-style techniques.

(2) Interpretation and use of the model requires domain expertise on the part of the analyst. Any effort to apply topic modeling to a corpus to answer interpretive questions must include a subject-area specialist on the team.

(3) Ultimately, the choice of models must be driven by the questions one asks. The process is empirically disciplined, in that, if the data are inappropriate for answering the analysts' questions, no topic model will produce a useful reduction of the data. And one can employ statistical tests, as we did, to ask if results meet the model's statistical assumptions and to learn from the deviations. But given a reasonable substantive fit, one may choose to trade off robustness for substantive interest in selecting a model.

(4) An important step in any analysis is to establish the model's validity. We employed three kinds of validity tests (Grimmer and Stewart, 2011; Mimno and Blei, 2011): statistical (the mutual information tests of solution fit to model assumptions); semantic (hand coding to determine whether the meaning of particular words varied significantly and as expected with assignment to different topics); and predictive (to see if events that should have increased the prevalence of particular topics if our interpretations are correct, actually did so).

(5) Our efforts at semantic validation exposed a limitation in the model's ability to discriminate among word senses for words assigned to topics with weak representation in a text. More research is needed to establish why this is the case; and new models that relax the bag-of-words assumption may solve this problem (Griffiths et al., 2005).

(6) More research is needed, as well, on how best to choose corpora and models to investigate heteroglossia and framing. Under what conditions will models yield topics that correspond to frames or voices (or both)? Developmental research will help us exploit the potential of this method to render operational these critical, but too rarely empirically analyzed, concepts.

(7) Finally, integrating topic models with affective analysis is an important priority. In interpreting our results, we contended that the shift in focus around 1989 toward the greater prevalence of topics dealing with controversy and contention (and away from those depicting arts grant-making in connection with positively valued cultural programs) produced a less positive discursive environment for government support for the arts. One possibility is to apply existing affective-analysis programs *ex post* to ask, for example, if the affective tone of references to arts agencies in texts dominated by Topics 2, 5, and 8 is more negative than in other cases. Another is to use affective-analysis programs to tag selected terms in advance in order to incorporate affective information into the model itself.

Topic modeling will not be a panacea for sociologists of culture. But it is a powerful tool for helping us understand and explore large archives of texts. Used properly by subject-area experts with appropriate validation, topic models can be valuable complements to other interpretive approaches, offering new ways to operationalize key concepts and to make sense of large textual corpora.

Acknowledgements

Research support from Princeton University and sabbatical support for the first author from the Russell Sage Foundation are gratefully acknowledged, as is the assistance of Brian

Steensland in gathering parts of the textual data. Support from the Rockefeller Foundation and the Andrew W. Mellon Foundation (through a grant to Princeton's Center for Arts and Cultural Policy Studies) for data collection, research and sabbatical support from Princeton University and the Russell Sage Foundation and research support from Princeton's Center for Information Technology and Public Policy are gratefully acknowledged. Thoughtful feedback from Amy Binder, Clayton Childress, Edward Hunter, the editor, special issue editor, and reviewers for *Poetics*, and participants in the University of Pennsylvania's Economic Sociology Colloquium and Princeton University's Theorodology Workshop is gratefully acknowledged.

Appendices. Supplementary information

Supplementary information associated with this article (Appendices A, B and C) can be found, in the online version, at http://dx.doi.org/10.1016/j.poetic.2013.08.004.

References

Alexander, J., 2000. Command Performance: An Actress in the Theatre of Politics. Public Affairs Press, New York.

Bakhtin, M.M., 1982 (1934–1941). (M. Holquist, Trans.) In: Emerson, C., Holquist, M. (Eds.), The Dialogic Imagination: Four Essays. University of Texas Press, Austin, TX.

Benford, R.D., 1997. An insider's critique of the social movement framing perspective. Sociological Inquiry 67, 409–430.

Benson, R., 2013. Shaping Immigration News: A French-American Comparison. Cambridge University Press, New York.

Bird, S.L., 2011. Seeking the audience for news: response, news talk, and everyday practices. In: Nightingale, V. (Ed.), The Handbook of Media Audiences. Wiley-Blackwell, Oxford, UK, pp. 489–508.

Blei, D.M., 2012. Probabilistic topic models. Communications of the ACM 5, 77–84.

Blei, D.M., Lafferty, J.D., 2009. Topic models. In: Srivastava, A.N., Sahami, M. (Eds.), Text Mining: Classification, Clustering, and Applications. Taylor and Francis, London, pp. 71–94.

Blei, D.M., McAuliffe, J., 2007. Supervised topic models. http://arxiv.org/pdf/1003.0783v1.pdf (accessed 23.08.13).

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

Boczkowski, P.J., 2010. News at Work: Imitation in an Age of Information Abundance. University of Chicago Press, Chicago.

Box, G.E.P., 1979. Robustness in the strategy of scientific model building. In: Launer, R.L., Wilkinson, G.N. (Eds.), Robustness in Statistics. Academic Press, New York, pp. 201–236.

Brenson, M., 1998. Washington's stake in the arts. New York Times April 12, http://www.nytimes.com/1998/04/12/arts/washington-s-stake-in-the-arts.html?src=pm (accessed 24.06.11).

Buis, M.L., 2008. FMLOGIT: Stata Module Fitting a Fractional Multinomial Logit Model by Quasi Maximum Likelihood. Economic Research Division, Federal Reserve Bank of St. Louis IDEAS Website; http://ideas.repec.org/c/boc/bocode/s456976.html (accessed 23.08.13).

Burgess, C., 2006. Multiple streams and policy community dynamics: the 1990 NEA Independent Commission. Journal of Arts Management, Law, and Society 36, 104–126.

Copestake, A., Briscoe, T., 1995. Semi-productive polysemy and sense extension. Journal of Semantics 12, 15–67.

DiMaggio, P., 1986. Introduction. In: DiMaggio, P. (Ed.), Nonprofit Enterprise in the Arts: Studies in Mission and Constraint. Oxford University Press, New York, pp. 3–14.

DiMaggio, P., 1997. Culture and cognition. Annual Review of Sociology 23, 263–287.

DiMaggio, P., 2003. The myth of culture war: the disparity between private opinion and public politics. In: Rieder, J. (Ed.), Fractious Nation: Unity and Division in Contemporary American Life. University of California Press, Berkeley, CA, pp. 79–97.

DiMaggio, P., Bryson, B., 2007. Public attitudes toward cultural authority and cultural diversity in higher education and the arts. In: Blake, C.N. (Ed.), The Arts of Democracy: Art, Public Culture and the State. University of Pennsylvania Press, Philadelphia, pp. 243–274.

DiMaggio, P., Cadge, W., Robinson, L., Steensland, B., 2001. The role of religion in public conflicts over the arts in the Philadelphia Area, 1965-1997. In: Arthurs, A., Wallach, G. (Eds.), Crossroads: Art and Religion in American Life. New Press, New York, pp. 103–138.

DiMaggio, P., Pettit, B., 1999. Public Opinion and Political Vulnerability: Why Has the National Endowment for the Arts Been Such an Attractive Target? Princeton University Center for Arts and Cultural Policy Studies, Princeton, NJ, Working Paper #7.

Doyle, G., Elkan, C., n.d. Financial Topic Models. http://www.umiacs.umd.edu/∼jbg/nips_tm_workshop/22.pdf (accessed 20.06.11).

Doyle, G., Elkin, C., 2009. Accounting for burstiness in topic models. In: Proceedings of International Conference on Machine Learning, Association for Computing Machinery, New York, pp. 201–208.

Dubin, S.C., 1992. Arresting Images: Impolitic Art and Uncivil Actions. Routledge, Boston.

Feldman, S., 2003. Enforcing social conformity: a theory of authoritarianism. Political Psychology 24, 41–74.

Fiss, O., 1991. State activism and state censorship. Yale Law Journal 100, 2087–2106.

Frohnmayer, J., 1993. Leaving Town Alive: Confessions of an Arts Warrior. Houghton Mifflin, Boston.

Gamson, W.A., 1992. Talking Politics. Cambridge University Press, New York.

Gamson, W.A., Croteau, D., Hoynes, W., Sasson, T., 1992. Media images and the social construction of reality. Annual Review of Sociology 18, 373–393.

Gelman, A., Meng, X., Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6, 733–807.

General Social Survey, 1998. Retrieved July 14, 2013 from the iPOLL Databank. The Roper Center for Public Opinion Research, University of Connecticut (USNORC.GSS98C.Q0693).

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis. CRC Press, New York.

Gentzko, M., Shapiro, J.M., 2007. What Drives Media Slant? Evidence from U.S. daily Newspapers. NBER Working Paper 12707. http://www.nber.org/papers/w112707.

Gerrish, S.M., Blei, D.M., 2010. A language-based approach to measuring scholarly impact. In: Proceedings of the 27th International Conference on Machine Learning. International Machine Learning Society, Haifa, Israel http://www.cs.princeton.edu/∼blei/papers/GerrishBlei2010.pdf (accessed 23.08.13).

Gerrish, S.M., Blei, D.M., 2011. Predicting legislative roll call votes from text. In: Proceedings of the 28th International Conference on Machine Learning. International Machine Learning Society, Bellevue, WA http://www.cs.princeton.edu/∼blei/papers/GerrishBlei2011.pdf (accessed 23.08.13).

Griffiths, T., Steyvers, M., 2004. Finding scientific topics. Proceedings of the National Academy of Sciences 101 (Suppl. 1) 5228–5235.

Griffiths, T., Steyvers, M., Blei, D., Tenenbaum, J., 2005. Integrating topics and syntax. Advances in Neural Information Processing Systems 17, 537–544.

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., Stewart, B.M., 2011. Text as Data: The promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. http://www.stanford.edu/∼jgrimmer/tad2.pdf.

Ho, D.E., Quinn, K.M., 2008. Measuring explicit political positions of media. Quarterly Journal of Political Science 3, 353–377.

Holsti, O.R., 1969. Content Analysis for the Social Sciences and Humanities. Addison-Wesley, Reading, MA.

Iyengar, S., Kinder, D.R., 1987. News that Matters: Agenda-Setting and Priming in a Television Age. University of Chicago Press, Chicago.

Iyengar, S., 1991. Is Anyone Responsible? How Television Frames Political Issues. University of Chicago Press, Chicago.

Janssen, S., Kuipers, G., Verboord, M., 2008. Cultural globalization and arts journalism: the international orientation of arts and culture coverage in Dutch, French, German and U.S. Newspapers, 1955–2005. American Sociological Review 73, 719–740.

Jensen, R., 1995. The culture wars, 1965–1995: a historian's map. Journal of Social History 29, 17–37.

Johnston, H., 1995. A methodology for frame analysis: from discourse to cognitive schemata. In: Johnston, H., Klandermans, B. (Eds.), Social Movements and Culture. University of Minnesota Press, Minneapolis, pp. 217–246.

Kimbis, T.P., 1997. Surviving the storm: how the National Endowment for the Arts restructured itself to serve a new constituency, Journal of Arts Management. Law and Society 27, 139–158.

Koch, C., 1998. The contest for American culture: a leadership case study on the NEA and NEH funding crisis. Public Talk: Online Journal of Discourse Leadership http://www.upenn.edu/pnc/ptkoch.html (accessed 14.06.12).

Klebanov, B.B., Diermeier, D., Beigman, E., 2008. Automatic annotation of semantic fields for political science research. Journal of Information Technology & Politics 5 (1) 95–120.

Kresse, M., 1991. Turmoil at the National Endowment for the Arts: can federally funded art survive the ''Mapplethorpe controversy? Buffalo Law Review 39, 231–273.

Krippendorf, K., 2004. Content Analysis: An Introduction to Its Methodology. Second edition. Sage, Thousand Oaks, CA.

McCombs, M., Shaw, D., 1972. The agenda setting function of the mass media. Public Opinion Quarterly 36, 176–187.
McLeod, D.M., MacKenzie, J.A., 1998. Print media and public reaction to the controversy over NEA funding for Robert Mapplethorpe's "The Perfect Moment" exhibit. Journalism and Mass Communication Quarterly 75, 278–291.
Mimno, D., Blei, D.M., 2011. Bayesian checking for topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Edinburgh, pp. 227–237.
Mohr, J., 1994. Soldiers, mothers, tramps and others: discourse roles in the 1907 New York Charity Directory. Poetics 22, 327–358.
Mohr, J.W., 1998. Measuring meaning structures. Annual Review of Sociology 24, 345–370.
Mohr, J.W., Duquenne, V., 1997. The duality of culture and structure: poverty relief in New York City, 1888–1917. Theory and Society 26, 305–356.
Molotch, H., Lester, M., 1974. News as purposive behavior: on the strategic use of routine events, accidents and scandals. American Sociological Review 39, 101–112.
Pettit, B., DiMaggio, P., 1997. Public Sentiments towards the Arts: An Analysis of 13 Opinion Surveys. Princeton University Center for Arts and Cultural Policy Studies, Princeton, NJ, Working Paper #5.
Price, V., Tewksbury, D., 1997. News values and public opinion: a theoretical account of media priming and framing. In: Barnett, G., Boster, F.J. (Eds.), Progress in Communication Science. Ables, Greenwich, CT, pp. 173–212.
Quinn, K.M., Monroe, B., Colaresi, M., Crespin, M., Radev, D., 2010. How to analyze political attention with minimal assumptions and costs. American Journal of Political Science 54, 209–228.
Ramage, D., Manning, C., McFarland, D., 2010.In: Which Universities Lead and Lag: Toward University Rankings Based on Scholarly Output. NIPS Workshop on Computational Social Science and the Wisdom of the Crowds. Whistler, Canada http://www.cs.umass.edu/∼wallach/workshops/nips2010css/papers/ramage.pdf (accessed 20.06.11).
Ramage, D., Rosen, E., Chuang, J., Manning, C.D., McFarland, D.A., 2009.In: Topic Modeling for the Social Sciences. NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond. Whistler, Canada http://www.umiacs.umd.edu/∼jbg/nips_tm_workshop/23.pdf (accessed 20.06.11).
Reese, S.D., 1991. Setting the media's agenda: a power balance perspective. Communication Yearbook 14, 309–340.
Rogers, E.M., Dearing, J.W., 1988. Agenda-setting research: where has it been? Where is it going?. Communications Yearbook 11, 555–594.
Rubin, D., 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Annals of Statistics 12, 1151–1172.
Saussure, F., 1983. Course in General Linguistics. Open Court Press, La Salle, IL.
Shaw, D., 1977. The press agenda in a community setting. In: Shaw, D., McCombs, M. (Eds.), The Emergence of American Public Issues. The Agenda-Setting Function of the Press, West, St. Paul, MN, pp. 33–51.
Snow, D.A., 2004. Framing processes, ideology and discursive fields. In: Snow, D.A., Soule, S.A., Kriesl, H. (Eds.), The Blackwell Companion to Social Movements. Blackwell Publishing, Chichester, UK, pp. 380–412.
Stone, P.J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M., 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge, MA.
Weaver, D.H., Mauro, J.B., 1978. Newspaper reading patterns. Journalism Quarterly 55, 84–91.
Ziegler, J.W., 1994. Arts in Crisis: The National Endowment for the Arts versus America. A Cappella, Chicago.

Paul DiMaggio is A. Barton Hepburn Professor of Sociology and Public Affairs at Princeton University. His interests include patterns of cultural participation, the organization of the Internet, formal and statistical methods of cultural analysis, and the impact of social networks on social inequality.

Manish Nag is a PhD candidate in the Department of Sociology at Princeton University. His research utilizes computational social science innovations in text analysis and social network analysis to understand cultural change in media and academic discourses, as well as change and resilience in global networks of people, goods and ideas.

David Blei is an associate professor of Computer Science at Princeton University. His research focuses on probabilistic topic models, Bayesian nonparametric methods, and approximate posterior inference. He works on a variety of applications, including text, images, music, social networks, and scientific data.