The effects of transitions of power on the contents of municipal government websites

Markus Neumann   Bruce Desmarais   Hanna Wallach

September 22, 2017

Abstract

We explore the effect of transitions of power in municipal governments on the content of their websites. We hypothesize that when party control changes, city administrators modify the contents of their websites in order to fit the agenda of the new incumbent. To test this theory, we study cities in Indiana and Louisiana, two states in which all municipal elections are partisan and the parties of the candidates appear on the ballots. Snapshots of websites before and after transitions of power are acquired through the Wayback Machine. We apply statistical topic models based in latent dirichlet allocation, focusing on changes to the websites. We present results on both which topics see the greatest degree of change associated with transitions in city administrations, and how the topics modified differ with regard to political parties.

1   Introduction

2   Background

Grimmelikhuijsen (2010) run an experiment in which citizens are exposed to randomly selected levels of information about local government council minutes. They find a negative relationship between the information level and perceptions of competence in the local government. This raises an interesting question regarding whether citizens or more likely to participate when they perceive competence or when they perceive incompetence.

Wang, Bretschneider and Gant (2005) present a widely cited 'model' for evaluating the accessibility of information on government websites. This is an important paper with which we should be familiar at a very detailed level as we use archived web content to assess the volume/accessibility of information provided by local governments.

Osman, Anouze, Irani, Al-Ayoubi, Lee, Balcı, Medeni and Weerakkody (2014) is less relevant, but they develop a multi-item measure to predict the level of citizen satisfaction with e-government services.

Grimmelikhuijsen and Welch (2012) conduct an enormously relevant study. Insofar as we analyze what predicts openness of government websites, we will be replicating and building upon this study. They focus on Dutch municipal websites, and their approach is fairly limited in scope and highly manual (which we can compliment). For example, one of the dependent variables "Decision-making transparency," is measured "using a discrete (1/0) indicator for whether the underlying principles or reasons for local air pollution policies were given on the Web site."

Names	Year	Journal	Findings	Important?
Benedictis-Kessner, Justin De Warshaw, Christopher	2016	JOP*	Regression discontinuity design. Democratic mayors spend more (but it is unclear on what, not the typical Democratic issue-areas), issue more debt, pay more interest	Yes
Caughey, Devin Warshaw, Christopher Xu, Yiqing	2015	Working Paper	Regression discontinuity design. Partisan composition of state governments affects state policy liberalism (composite index for the areas of social welfare, taxation, labor, civil rights, women's rights, moral legislation, family planning, environment).	Somewhat
Einstein, Katherine Levine Kogan, Vladimir	2015	Urban Affairs Review	Cities with more Democratic citizens spend more; more progressive (rather than regressive) forms of taxation; pursue intergov. aid more; spend more on police, fire, parks & recreation	Somewhat
Einstein, Katherine Levine Glick, David M.	2015	Working Paper	Survey of 72 mayors. Unlike Republican mayors, roughly half of Democrats seem to agree that cities should aim to reduce inequality. Democratic mayors also seem to favor redistribution to accomplish that goal.	Somewhat
Kiewiet, D Roderick Mccubbins, Mathew D	2014	Annual Review	City budgets have been severeley constrained since the Great Recession. Spending has thus decreased in general. Lack of funds means that there is not much discretion for partisanship.	Somewhat
Tausanovitch, Chris Warshaw, Christopher	2014	APSR*	Cities are responsive (taxes, expenditures, regressiveness of taxation) to citizens' conservatism/liberalism. Partisan elections do not make cities more or less responsive.	Yes
Guillamón, Ma Dolores Bastida, Francisco Benito, Bernardino	2013	European Journal of Law and Economics	Police spending in Spain. Conservative parties spend more on police. Spending is higher before elections. Also contains a useful overview of the literature.	Yes

Names	Year	Journal	Findings	Important?
Gerber, Elisabeth R.	2013	Cityscape	Partisanship of both citizens and elected city officials separately affect climate policy.	Yes
Solé-Ollé, Albert Viladecans-Marsal, Elisabet	2013	Journal of Urban Economics	Spanish cities. The authors "employ a regression discontinuity design to document that cities controlled by left-wing parties convert much less land from rural to urban uses than is the case in similar cities con- trolled by the right". Partisanship might also affect housing construction and price growth.	Yes
Gerber, Elisabeth R. Hopkins, Daniel J.	2011	AJPS	Regression discontinuity design. Democratic mayors spend less on public safety. All other policy areas (including taxation) are unaffected.	Yes
Trounstine, Jessica	2010	Annual Review	Race and ethnicity in local elections (not relevant to us). Partisan elections have higher turnout; non-partisan elections still tend to have some partisanship in them because voters learn about party of candidates from media. Non-partisan elections favor Republicans/upper class. Mixed evidence for whether partisanship of mayor is important for policy.	Somewhat
Palus, Christine Kelleher	2010	State and Local Government Review	Ideology (liberal/conservative) of citizen is well represented by gov. spending in five areas: (1) community development, housing, and conservation, (2) health and human services, (3) culture, the arts, and recreation, (4) environmental programs, and (5) transportation.	Somewhat
Ferreira, Fernando Gyourko, Joseph	2009	The Quarterly Journal of Economics	Regression discontinuity design. Null results for spending and city gov. size with regard to mayor partisanship.	Yes
Ansolabehere, Stephen Snyder, James M.	2006	Scandinavian Journal of Economics	Despite the journal, this is about the U.S. The important finding (for us) is the fact that counties whose government is controlled by the same party as the state government, receive more funding (county's share of state transfers, normalized by county pop.) from the state.	Somewhat
Murphy, Russell D.	2002	Annual Review	Not useful. Too philosophical; mostly cites papers written a hundred years ago. Also exclusively about larger cities.	No

3

Names	Year	Journal	Findings	Important?
Armstrong, Cory L.	2011	Government Information Quarterly	Comparison of county and school board websites in Florida (where the two align) with regard to transparency (presence or absence of public records). Manual content analysis (undergrads told to look around for 15 minutes). School board websites, more professional websites, and websites in Republican-dominated counties are found to be more transparent.	Yes
Cegarra-Navarro, Juan Pachón, José Cegarra, José	2012	International Journal of Information Management	Survey of Spanish municipal government officials (specifically, the city website managers). Respondents are asked about the features of their websites, the level of civic engagement and the size of their municipality. More sophisticated websites are correlated with greater civic engagement and greater use of e-government functions.	Yes
Dolson, Jordan Young, Robert	2012	Canadian Journal of Urban Research	Determinants of website content. Three categories: e-content (city information on website), e-participation, social media use. Tables on page 15 show frequencies of these categories across sites, and might be useful to inform our topics. Larger cities have better websites. Population growth and immigration are also tested, but the findings are somewhat inconclusive.	Yes
Feeney, Mary K. Brown, Adrian	2017	Government Information Quarterly	500 U.S. city websites at two points in time (2010-2014). Count model of website features regarding information, e-services, utilities, transparency and civic engagement. Having a larger population leads to more features. Relying on a website contractor leads to more information and transparency. The authors say that mayor-councils are negatively correlated with website sophistication, but their regression tables state the opposite.	Yes
Kaylor, Charles Deshazo, Randy Van Eck, David	2001	Government Information Quarterly	Model of best practices of e-government. Table 1 lists a number of possible ways this manifests, could be useful for our theory.	Somewhat
Ansolabehere, Stephen Urban, Florian	2002	Cities	Websites of 20 major cities across the world. Is website content correlated with city characteristics? Not particularly systematic, and the findings are inconclusive.	Somewhat
Jeffres, Leo W. Lin, Carolyn A.	2006	Journal of Computer-Mediated Communication	50 largest metropolitan areas in the U.S. Features include information about city, opportunities for citizen feedback, galleries of photos, links, etc. Purely descriptive analysis, doesn't contain anything that isn't covered in any of the other aricles.	No

There doesn't seem to be much literature on transitions of party control, if anything, that question is mostly phrased with regard to political representation. However, if we want to tie our paper to a larger theory, we could go with dynamic representation. Under dynamic representation, policy-makers are responsive to trends in citizens opinions, which mainly manifest/become apparent through election results, especially when incumbents are voted out of office. This fits our topic quite well. Also, of the few papers that do exist on the effects of partisan transitions, virtually all use regression discontinuity designs.

Why does the content of city websites matter? According to Mayhew (1974), politicians engage in advertising, credit claiming and position taking in order to get re-elected. Official city websites allow mayors to do all three. Their offices frequently take a prominent position on the frontpage, and many websites also feature a picture of the candidate. In local politics, where campaign funds are low, this lends the incumbent a crucial advantage in becoming more well-known among her constituents. Furthermore, municipal politics gives incumbents clear and tangible achievements they can point to, such as completed infrastructure projects, the acquisition of federal or state funding, or the hosting of city-wide events. City websites present an opportunity for local officials to brandish these accomplishments. Finally, they also give mayors a platform from which they can advertise their political beliefs. On municipal websites, this may not manifest in the form of brazen partisanship, but more subtle avenues are available. As noted by Einstein and Glick (2015), there are stark differences in the spending preferences of Democratic and Republican mayors. City websites can then be used to communicate the stance of a mayor on social or economic programs. Another advantage of websites with regard to communication is that unlike direct social interactions, officials have full control over them.

In addition to the use of city websites for the politicians that control them, variance in content also matters with regard to the people who visit them. Local residents likely rely on city websites to get news about events, hot-button political issues specific to their city, contact city officials or find out addresses or opening hours of city institutions. Visitors use city websites to look up local attractions, which are often described in great detail. Similarly, prospective residents looking to move, might rely on city websites to inform their decision on whether to relocate there. An inviting website emphasizing the city's receptiveness to new residents might make a real difference here. Finally, city websites frequently feature sections on business, but there is a lot of variance in this area: Some emphasize economic development, properties, or transportation, whereas others focus on undeveloped land and other business opportunities. Differences in websites likely say something about a city's economic profile, with potential repercussions for the political realm.

3 Data

The General Services Administration (GSA) maintains all .gov addresses, and provides a complete[1] list of all such domains to the public through GitHub[2]. This list is updated once per month -

---

[1]Domains used for testing and internal programs are excluded.
[2]https://github.com/GSA/data/tree/gh-pages/dotgov-domains

we rely on the version released on January 16, 2017. The data from the GSA contains the following variables: One, domain name, specifically, the all-uppercase version of domain and top-level domain (for example, 'ABERDEENMD.GOV'). Two, the type of government entity to which the domain is registered, such as city, county, federal agency, etc. Three, for federal agencies, the name is specfied. Finally, the city in which the domain is registered, is noted.

Here, we focus only on cities. As a first step, we use a webdriver-controlled browser (Firefox/Selenium/Geckodriver) to test whether all of the city websites actually work. Of the 2425 domains listed by the GSA as cities, 292 are not accessible. Furthermore, the .gov domain, as registered at the GSA, is frequently not the website a city actually uses. In many cases, these sites redirect to another address, sometimes not a .gov domain (in this case, we simply use this domain). We record these URLs, as they are required to retrieve the images websites stored in the Wayback Machine (WbM).

$T_1 + \beta_1$

In order to provide an overview of our coverage (as not all cities, towns and villages use .gov addresses), we merge this list with U.S. Census data[3]. Here, several limitations in the GSA data need to be accounted for: One, even though the GSA nominally separates websites of cities and counties, some of the domains categorized as cities actually belong to counties. The same is true for townships and boroughs. Ergo, we eliminate all websites belonging to these three types of entities by hand. Furthermore, the city name, as given by the GSA, refers to the city in which the domain is registered, which is not necessarily equivalent to the city the website serves. In many cases, a website of a larger city may be registered to one of its subdivisions (for example, the website of New York is registered to Brooklyn), or vice versa (for example, the website of Homecroftin, a small town within Indianapolis, is registered to the city as a whole). Consequently we fix mismatches between websites and cities manually. Finally, a number of cities are simply misspelled, which we also correct by hand.

After the counties, townships and cities that cannot be matched to the Census data[4] and duplicate websites (some cities have more than one website) are removed, 1813 domains/cities remain.

These cities contain 90,616,865 people, and thus about 28% of the U.S. population (see figure 1).

We use the resulting list of websites to acccess their copies stored in the Internet Archive's Wayback Machine. To this end, we rely on the Ruby Gem 'Wayback Machine Downloader'[5] (WbMD). We supply the URL that each .gov website redirects to to the WbMD, which then downloads every file present in the WbM from a snapshot in October 2016, or, if not available, as soon as possible after this point.

<Note: We have not actually done this last step for all websites (however, the R script which runs the Ruby package is already set up to do so once we need to). Instead 10 websites were randomly sampled from an older version of the GSA list, which still contained counties and townships, which is why one of the 10 websites is from Dutchess County, NY.>

---

[3]http://www2.census.gov/programs-surveys/popest/datasets/2010-2015/cities/totals/sub-est2015_all.csv
[4]There are five cities that are not contained in the Census data
[5]https://github.com/hartator/wayback-machine-downloader

Figure 1: Percentage of state population covered.
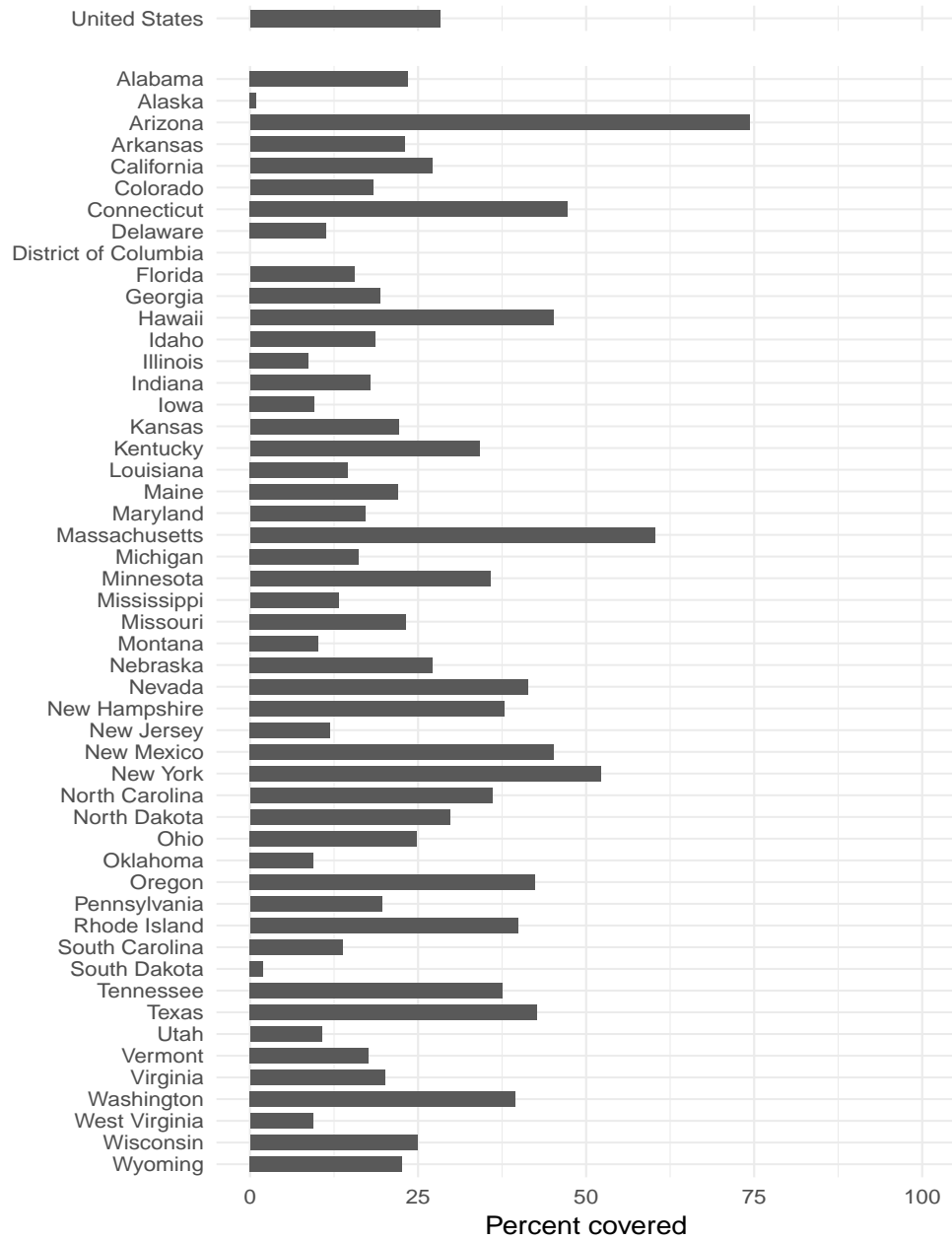
Filetype	current	before	after
	51455	13866	19199
pdf	9646	5489	7544
jpg	5216	1988	3512
html	3767	17842	17596
aspx	2832	4356	3271
png	2714	2327	3684
gif	1068	664	1077
JPG	478	182	263
1	443	61	54
css	390	265	518
js	350	255	468
htm	264	295	256
docx	203	106	120
doc	167	70	130
asp	161	201	211
svg	87	55	69
php	83	157	241

Table 1: The most common file types in scraped websites

Website	current_size	current_files	before_size	before_files	after_size	after_files	size_change	files_change	control_change
attica-in.gov	61988	1417	7528	164	55956	1390	7.43	8.48	0.00
bedford.in.us	57628	560	27452	182	46388	525	1.69	2.88	0.00
cityofboonvilleindiana.com	9848	110	16996	172	20784	229	1.22	1.33	0.00
frankfort-in.gov	205368	2652	12208	242	138360	1077	11.33	4.45	0.00
warsaw.in.gov	298440	2117	26844	539	360400	2036	13.43	3.78	0.00
www.bloomington.in.gov	131128	2713	443360	14384	247096	9640	0.56	0.67	0.00
www.brazil.in.gov	43056	845	34472	625	55152	1214	1.60	1.94	0.00
www.carmel.in.gov	2270016	8727	1919344	5361	899900	2219	0.47	0.41	0.00
www.ci.auburn.in.us	183296	1025	21444	345	23564	211	1.10	0.61	0.00
www.cityoffortwayne.org	2136424	4378	266784	3582	233600	3018	0.88	0.84	0.00
www.cityofhobart.org	722000	2463	44192	650	62660	1037	1.42	1.60	0.00
www.evansvillegov.org	6345932	11844	290784	1281	1697224	6853	5.84	5.35	0.00
www.gary.in.us	373888	1227	121812	485	157140	719	1.29	1.48	0.00
www.huntingburg-in.gov	388680	2496	8644	213	375900	1953	43.49	9.17	0.00
www.jasperindiana.gov	561968	4013	55900	460	439072	2224	7.85	4.83	0.00
www.lakestation-in.gov	48	2	7724	84	257272	1097	33.31	13.06	0.00
www.linton-in.gov	32	1	24	2	24	2	1.00	1.00	0.00
www.madison-in.gov	531044	1848	36636	575	191624	1444	5.23	2.51	0.00
www.martinsville.in.gov	46792	1463	71628	1052	80944	800	1.13	0.76	0.00
www.monticelloin.gov	33656	753	18120	448	100680	2104	5.56	4.70	0.00
www.newhavenin.org	84364	626	2524	86	6792	334	2.69	3.88	0.00
www.richmondindiana.gov	250968	1042	217252	918	401672	2422	1.85	2.64	0.00
www.southbendin.gov	1264076	4749	454456	3286	1424136	2562	3.13	0.78	0.00
connersvillecommunity.com	170688	569	162316	815	187276	808	1.15	0.99	1.00
www.batesvilleindiana.us	166564	2348	39592	496	95696	1310	2.42	2.64	1.00
www.cityofrisingsun.com	994956	3311	321400	1268	80848	868	0.25	0.68	1.00
www.cityofrockport-in.gov	12068	98	5148	16	12068	98	2.34	6.12	1.00
www.elkhartindiana.org	1132828	2345	5588	123	6204	223	1.11	1.81	1.00
www.elwoodcity-in.org	224412	765	5000	123	139692	517	27.94	4.20	1.00
www.indy.gov	5726048	9675	6119260	10451	4984080	7981	0.81	0.76	1.00
www.northvernon-in.gov	272016	403	3132	112	289336	416	92.38	3.71	1.00
www.winchester-in.gov	364592	2480	6508	135	45488	567	6.99	4.20	1.00

Table 2: Number of files and size of websites

For some cities, whose websites make heavy use of JavaScript, this method does not lead to satisfying results. Consequently we restricted our corpus to cities with at least 3 documents.

4   Preprocessing

The documents are read in line by line, converted to UTF-8 and then stripped of dates, punctuation, numbers and words connected by underscores. At this point, the documents of one city still closely resemble one another in the form of boilerplate content, be it website elements (i.e. "You are here", "Home", "Directory" etc.) in html documents, or commonly used forms or phrases in pdfs, doc and docx files. This is an issue, because it clusters documents around the cities from which they originate in a way that has nothing to do with their actual content. In other words, the signal would be drowned out by the noise. Consequently we remove this content as following: Each line of every document is compared to every line in every other document belonging to the same city. We count how many times each line is duplicated for that city. We remove any line occuring more than our chosen threshold of 10.[6] This means that each document only retains the information that is particular about it. Preprocessing further includes setting every character to lowercase, as well as the removal of bullet points which frequently occur in html documents, extraneous whitespace, xml documents mislabeled as html files, and empty documents. Furthermore, some documents contain gibberish, often as a result of faulty or impartial OCR. To combat this problem, we employ two solutions. One, we use spellchecking, implemented through the hunspell R package, to remove all non-English words. However, hunspell does not cover everything, either because some tokens are not actual words (for example artifacts from defective encoding), or because random sequences of characters just so happen to form words that exist in a dictionary (for example "eh" or "duh"). Since we rely on a bag-of-words model in which syntax does not matter, we can ameliorate these problems by removing all text except for whitespaces and the characters that appear in the english alphabet. Since a lot of the nonsensical text tends to be quite repetitive, we also delete all documents in which the proportion of unique to total number of tokens is less than 0.15. Furthermore, hunspell does not spellcheck individual characters, so we remove all individual characters appearing as tokens except for "i" and "a". Since these pre-processing steps reduce documents which are largely unsuitable to only a few words of texts that don't make much sense, we also remove all remaining documents containing less than 50 tokens.

4.1   Indiana City Websites

It would be fine to focus on Indiana as a case. First, we need to answer some preliminary questions about the data.

1. For what percentage and number of IN cities can we find data from the WBM?

2. For how many election cycles can we find political leadership data for these matched cities?

---

[6]Empirically, lines tend to be duplicated either hundreds of times, or only once or twice, if at all.

3. In what number and percentage of cities is the local leadership majority Republican?

4. Relatedly, in a typical election cycle, for how many cities do we see a transition in party leadership (i.e., a shift from majority D (R) to majority (R) D).

1. 30 cities, with a combined population of 1,180,435. However, since only cities (as opposed to towns and villages) hold mayoral elections, only 16 of these, with a combined population of 1,094,383 can be matched to the election data.

2. 2015, 2011, 2007, 2003.

3. Of the 16 cities, 7 have Republican mayors after the 2015 elections.

4. In 6 cases, a shift of party control occurs, with 4 of these being Republican –> Democratic.

4.2   Research Design

Variable	Unit	Source
Population size	1000 people	Census
Population growth last 5 years	Percent	Census
Type of economy (agriculture/industry/services)	?	Census
Economic performance (GDP?)	$	Census
Party of mayor before election	Rep/Dem/(Ind)	in.gov/sos/elections/
Party of mayor after election	Rep/Dem/(Ind)	in.gov/sos/elections/
Change of party control	0/1	in.gov/sos/elections/
Presidential vote 2012 in county	Percent Rep	? (but I have the data)
Unemployment rate	Percent	Census
Broadband speed	Avg. Mbps DL	broadbandmap.gov

Table 3: List of covariates

1. Corpus:

    (a) Last snapshots before the election (November 3, 2015 in Indiana; tbd. in Louisiana (probably February))

    (b) First snapshot that is at least 2 months after the new government's inauguration (which is in January for Indiana, May for Louisiana)

2. Preprocessing:

    (a) restrict corpus to:

        i. documents belonging to cities in which a change of power occurred

> ii. documents that were added, deleted or changed between the two snapshots

   (b) words to lowercase

   (c) remove punctuation

   (d) stemming (Porter stemming algorithm?)

   (e) Remove stop words (regular list of stop words is enough, since we use an asymmetric prior)

3. Apply Grimmer's expressed agenda model to the corpus

   (a) Asymmetric prior

   (b) Each document can have only one topic (in contrast to the author-topic model)

   (c) Cities $i = 1, ..., n = 15$

   (d) Topic $k(k = 1, ..., K)$

   (e) Documents $j(j = 1, ..., D_i)$ from city i

   (f) Party covariate in the prior, where the deleted and unmodified documents are coded as from the first, and the added and modified documents from the second party

4. Results

   (a) Label topics using Grimmer's automatic cluster labeling method, based on most commonly used words in documents belonging to topic

   (b) Evaluate topics

Validation:

- Do the above for cities in which no change of power occurred.

- Check whether there is higher than average turnover around the new year by comparing changes to non-election years (and also Louisiana, where elections are later).

- Check how long documents stay on websites on average. Use websites with a lot of snapshots for this (these exist for both small and large cities).

Problem with using this model: Grimmer's expressed agenda model uses Senators as the actors. Senators is also who he is substantively interested in. For us, the equivalent to Senators is cities. However, we care about parties, not cities.

4.3 Survival model

The existence of individual documents on municipal government websites can be though of as a survival process. No document stays on a website forever, and it appears to be a reasonable assumption that as documents get older and thus less relevant, they get replaced. The factors determining the steepness of the survival curve are the topic - fire safety regulations likely stay up longer than a bulletin on the annual spring banquet - and the change of party control after an election.

H1: The older a document, the more likely it is to be removed.

$S(t)$ has a downward slope. Admittedly, this is almost impossible not to be true. Also, test proportional, rising and falling hazard models.

H2: Documents pertaining to administrative matters are less likely to be removed.

Introduce a categorical variable for the top 10(?) topics. A negative coefficient for administrative topics would support this hypothesis.

H3: Documents introduced by the opposing party are more likely to be removed.

Introduce two variables into the survival model: One variable indicating which party has introduced a document, and a time-varying variable describing which party is currently in government. The hypothesis is tested through an interaction term between the two.

H4a: Democrats are more likely to remove documents with topics pertaining to private enterprise, private schools.

Interaction term between party in power and categorical topic variable.

H4b: Republicans are more likely to remove documents with topics pertaining to social justice, equality, taxes, public schools, etc.

Interaction term between party in power and categorical topic variable.

H5: In line with their commitment to small government, Republicans are more likely than Democrats to remove documents.

Party in power variable.

This model will take up a lot of degrees of freedom. The rarity of snapshots for some cities might be a problem. Documents being changed and being removed can be modeled as competing risks.

$$
\begin{aligned}
Y = {} & \text{Party that introduced the document} \\
& + \text{Party that is currently in power} \\
& + \text{Topic 1, topic 2, ..., topic k} \\
& + \text{Party that is currently in power} \times \text{Topic 1, topic 2, ..., topic k} \\
& + \text{Days since start of mayoral term (control)}
\end{aligned}
\tag{1}
$$

4.4 Topic modeling

Note: this chapter is mostly a wordy and less coherent version of the above.

We hypothesize that a change in leadership from one party to the other will lead to a change in website content because the two parties have different agendas. Democrats have a predilection towards policies that promote social and economic equality, whereas Republicans like to emphasize small government as well as law and order. Documents uploaded to city websites are expected to be a reflection of these preferences.

The Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan 2003) is the most commonly used topic model. However, it is unable to account for the existence of two parties with very different policy agendas, translating to different preferred topics. There are two types of extensions to the LDA that fit our subject much better - the structural topic model, and the author-topic model.

The structural topic model, developed by Roberts, Stewart and Airoldi (2016) allows researchers to model a corpus as a function of metadata associated with its documents. Specifically, topic prevalence (the proportion of a document made up by a topic) and topical content (the rate at which words figure into a topic) are contingent on a set of covariates. In our case, the two most important covariates are (1) city and (2) authoring party (operationalized by whether a document was present before a change of power, or introduced afterwards). Furthermore, the population size of the city should be a predictor for both the number and kind of problems it faces, which thus need to be addressed on its website. Furthermore, city size also serves as an estimate for the budget and technical capacities of its staff in charge of maintaining the website[7]. Further demographic as well as economic data might also be useful to differentiate cities from one another. If we model the differences between cities properly, we might not have to/should not include city as a (categorical) variable, because it would probably interfere with these more meaningful covariates.

The author-topic model (Rosen-Zvi, Griffiths, Steyvers and Smyth 2004) would allow us to capture the fact that different authors have different topical preferences. Unfortunately, we have two types of 'authors' - cities, and parties. Given the largely divergent administrative needs of different types of cities, we would likely have to treat cities as the author. This would require us to capture the partisan authorship of different documents entirely on the basis of sub-sets of the website data - changed, added, or deleted documents. (Note: In the papers on author-topic models,

---

[7]Although this relationship is not exactly deterministic - when looking through .gov websites manually, I've noticed that a lot of websites of (presumably wealthy) towns of only a few thousand citizens often have extremely well-kept websites

the intention is often to analyze scientific articles. These articles are often co-authored. Would it be possible to have BOTH cities and parties as authors, so that a specific version of a website would then be 'co-authored' by its city and party?)

The critical element in this analysis is to accurately attribute authorship of documents to either party. Despite possible changes to websites due to a leadership transition, large parts of the content carry over. This means that unless the successor government decides to delete everything, some of the existing documents will be preserved, and in the model, also attributed to the new 'author'. But the reverse is not possible, because the predecessor government can't choose to retain documents from the future. *This is a very important point for municipal websites. We should investigate the possibility of modeling only the changes—documents that change, documents that are deleted, and documents that are added.*

Labeling newly added documents after a change of power is quite simple. As far as older documents are concerned, we would have to operate under the assumption that the incumbent didn't keep his or her successor's document's on the website for four years.[8] One problem here is the fact that the incumbent would have all the administrative topics assigned to them, simply because they have to have those on their website.

If we really do end up getting swamped with administrative terms in our topic models (and it does kind of look like that at the moment), we might be able to separate the signal from the noise by running a preparatory LDA once and using its results to create a new, corpus-specific list of stop words. After that, we run the actual model. This way, politically charged terms and topics, which likely are not as common, but present nevertheless, should be able to rise to the surface. It might be possible to refine this process by running an exploratory model on website data from cities in which party control never changes, and the incumbent always wins by large margins. 'Safe' cities like this should have fairly homogeneous populations, with little need for the incumbent to play politics on the municipal website. Hence, these websites should be filled with purely administrative content.

The use of asymmetric priors (Wallach, Mimno and Mccallum 2009) over the document topic distribution - i.e. the assumption that some topics, such as administrative content, are inherently more common - may be a more elegant way of dealing with this issue.

Another intervening factor is that for cities in Indiana, mayoral terms begin in January. Since a lot of clerical and administrative tasks tend to be year-specific, work tends to pile up around the new year. Thus it is possible that a spike in newly added documents is not due to a change in party control, but owed to a seasonal increase in activity. We can test for this by comparing election years to non-election years. Furthermore, since in Louisiana, mayors take office in May, we have another point of comparison.

Furthermore, if we only investigate cities in which control of government changes from one party to another, we may overestimate its effect. Not only does a transition in party control occur,

---

[8]Probably a safe assumption. However, we could, and probably should test how long documents tend to stay on a city website. Simple descriptive statistics (for example density plots) on the length of existance should likely be sufficient. If we want to be really fancy about it, we could create a duration model, with document topics as features. This would allow us to measure whether some documents tend to remain longer based on their topic (i.e. fire regulations are probably going to stay up longer than notes on a specific council meeting).

but the person in charge also changes. Parties are fairly homogeneous, so that two mayors from the same party may have very different policy preferences and managerial styles. To remedy this problem, we [could] utilize matching, pairing our cases with similar cities in which the incumbent does not run for re-election, but party control stays the same nevertheless.

In order to determine the optimal number of topics for our corpus, we utilize the measure developed by **?**. In contrast to other approaches such as **?** or **?**, this measure includes as much information about the model as possible by taking into account both the document-term matrix, as well as the word-topic matrix. We find 200 topics to be the best value (see figure 12).

4.4.1   LDA diagnostics

Figure 3 plots the densities of topic weights across documents, where each line represents a topic. Distributions with lower peaks near zero and flatter bodies indicate a topic that is present to varying degrees in multiple documents. This shape appears to be more common in documents from Republican cities. By contrast, a distribution with a high peak near zero and a very long tail is indicative of a topic that only appears in one specific document. This appears to be more common for documents in Democratic cities. The substantive conclusion then is that Democrats appear to be more prone to dedicating individual documents to one specific purpose, whereas Republicans produce general-purpose documents more often.

An alternative way to consider topic weights is to aggregate them across documents, in this case through their median or mean. Figure 4 shows this distribution of topic weights across topics. Most topics have a very low weight - meaning that they do not appear in a lot of documents. When aggregating across documents via the mean, this effect is more pronounced for Republicans, suggesting that most topics do not feature frequently (or even at all) in their documents. Democrats on the other hand appear to have a wider spectrum of topics from which they chose - supporting the hypothesis of Democrats as a 'big tent' party. However, this effect is reversed when using the median instead of the mean to aggregate across documents. Now, the the distribution has a higher peak and lower tail for Democrats. The cause for this stark divergence appears to lie in the fact the mean and median differ enormously - for the range displayed here, the mean is about a hundred times greater. Evidently, extremely large weights (i.e. some documents fitting specific topics perfectly) distort the picture. The median document is a better representation of the corpus as a whole, but the results do not fit our hypothesis. Their are however more consistent with figure 4.

Topics for which there are stark differences in their distribution across documents between Democrats and Republicans are of particular interest to us. To detect these topics, we take the absolute difference of the median document topic weights, and arrange them as a histogram, see figure 5. For the most part, differences are small. For a few topics however, a contrasts emerges.

Figure 6 displays the word-topic probabilities for the 10 topics with the largest partisan differences. Some of these, such as topic 82, with its focus on policing, safety and crime could be construed as politically charged. For the most part however, these topics focus on administrative matters.

16

Figure 2: Dates of Wayback Machine snapshots. The vertical lines are municipal elections.
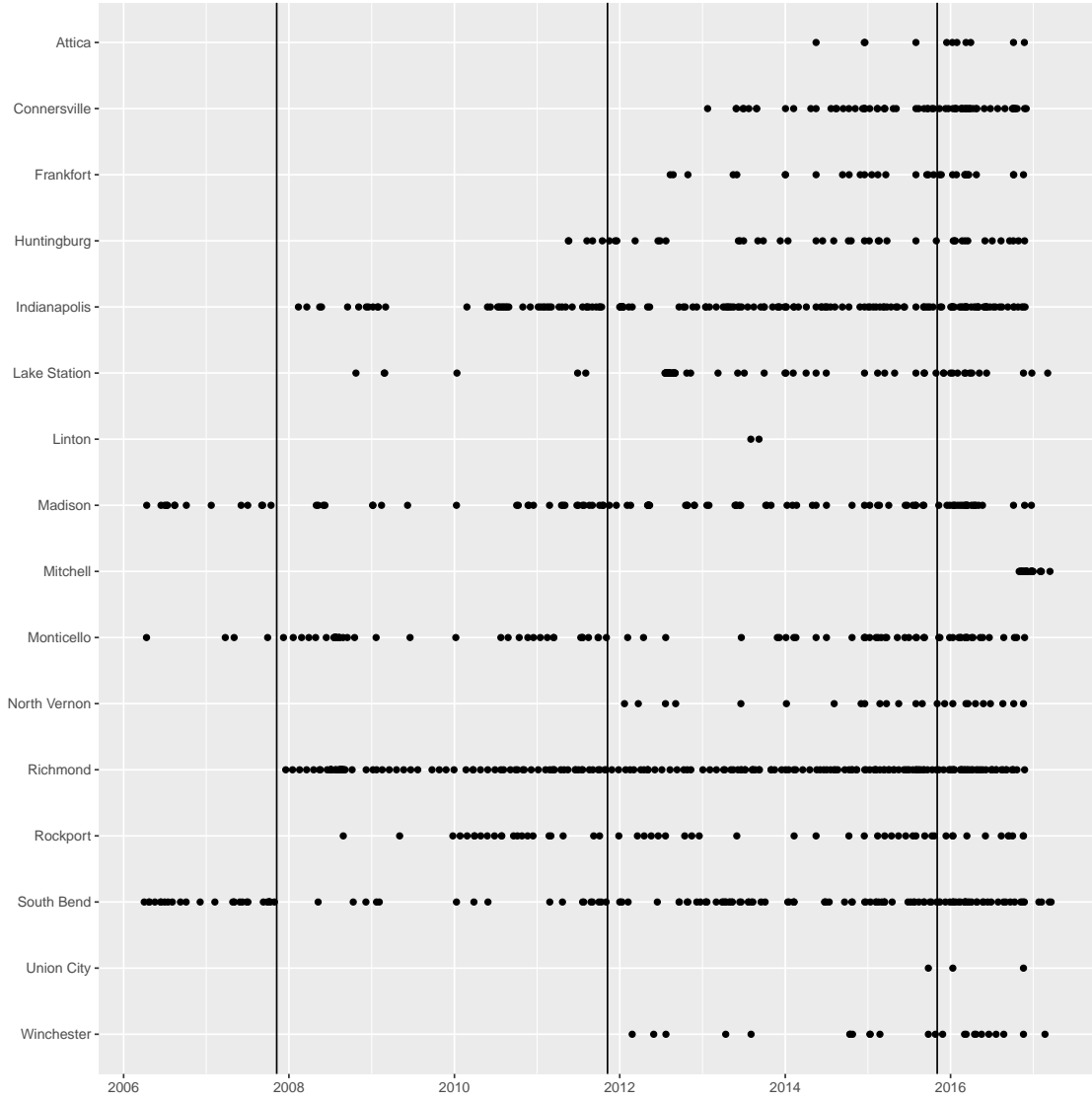
To further investigate the topics with large partisan differences, figure 6 shows densities of topic weights just like figure 3, but reduced to only the topics with the largest differences. It appears that mere absolute differences between the medians (or some other measure of central tendency) obscures the fact that the distributions themselves are quite different.

To further investigate this issue, we look at the raw data itself - the document-topic matrices for Democrats and Republicans, displayed as heatmaps. Figure 8 shows that especially for Re-

publicans (but also for Democrats, to a lesser extent), there appears to be extensive clustering for consecutive documents. Since the order of the documents in the corpus is dependent on the cities in which they appear in, it seems that topics are mere representation of city websites - each website 'owns' a number of topics, that appear across all of its documents, and hardly anywhere else. One possible cause for this type of clustering is the fact that documents frequently share common words, for example pertaining to navigation on the site, or standard forms that are shared throughout all documents.

4.4.2   Structural topic model

An ostensibly intuitive solution to topics clustering into cities in LDA is to include dummies for the cities in a statistical model of topics. This is facilitated by the structural topic model, which uses metadata on the document to account for variation in topics (Roberts, Stewart, Tingley, Lucas, Leder-Luis, Gadarian, Albertson and Rand 2014). However, figure 13 shows that if anything, the STM exacerbates the problem. Here, we plot the *p-values* of the coefficients for each city as well as the party variable across each topic. Under normal circumstances, plotting the p-values, as opposed to the fitted values, does not make much sense, but here it serves a diagnostic purpose. The plot shows that the party variable is never statistically significant at any conceivable level of confidence, nor is it even close to. Interestingly the same is true for a number of the cities as well. The topics cluster heavily into only about half of the cities, which does not present an improvement over LDA at all.

4.4.3   Prediction with SVM

An alternative approach to the problem is to ignore topics entirely, and go straight to predicting documents that are much more likely to be included on websites belonging to one or the other party. Classic machine learning techniques such as Naive Bayes, Linear Discriminant Analysis, or SVM should be expected to fare well in this context. Here, we rely on SVM, through three different implementations. The SciKitLearn package in Python, the e1071 package as well as the mlr package (relying on the kernlab package) in R. Predicting based on the document-term matrix (either in its regular form, or as its tf-idf representation), and using five-fold cros-validation, SciKitLearn produces a predictive accuracy of at most 78%. Parameter tuning did not lead to any improvements. The e1071 package in R, relying on the top 5000 features ranked through tf-idf fared slightly better, at 82 percent correctly predicted. Finally, the mlr/kernlab implementation (as well as the e1071 implementation through mlr) has not worked at all so far, resulting in stack overflow errors (which appears to be due to memory limitations in R).

Another point worth noting is that Monroe, Colaresi and Quinn (2008) advise against using these types of methods in this context because they get the data generation process backwards: Our theory assumes that party leads to variation in writing, and yet we rely on the documents to predict party, in spite of the fact that we actually have perfect knowledge of it.

Other methods used: Elastic-net in the glmnet package in R. Accuracy is 0.6924795 for in-sample prediction, so not worth bothering with.

dem.word	dem.tokens	rep.word	rep.tokens
city	52829	shall	28205
said	40549	street	26583
county	37440	city	25874
proposal	29068	plan	18971
public	26768	board	17875
department	26061	water	17753
services	23144	public	14629
fund	21809	council	13731
property	19739	use	12943
stated	19054	new	12367
council	18859	project	12139
community	18812	area	11199
development	18027	information	11028
budget	16582	meeting	10955
park	16478	fire	10950
asked	16464	ordinance	10728
board	15984	must	10421
tax	15857	motion	10151
can	14429	building	10070
office	14240	section	9689
information	13920	construction	8813
new	13639	made	8487
state	13462	department	8419
program	13310	request	8321
also	13122	service	8297
service	13055	ave	8287
one	12862	one	8089
system	12854	per	8037
project	12532	road	7881
provide	12292	year	7868

	Democratic	Republican
Cities	15	17
Documents	10257	5859
Tokens	10091324	3701293
Token assignments	5953078	2238028
Topics	101	99

References

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Einstein, Katherine Levine and David M. Glick. 2015. "Mayors, Partisanship, and Redistribution: Evidence Directly From U.S. Mayors.".

Grimmelikhuijsen, Stephan G. 2010. "Transparency of Public Decision-Making: Towards Trust in Local Government?" *Policy & Internet* 2(1):5–35.

Grimmelikhuijsen, Stephan G and Eric W Welch. 2012. "Developing and testing a theoretical framework for computer-mediated transparency of local governments." *Public administration review* 72(4):562–571.

Mayhew, David. 1974. *Congress: The Electoral Connection*. Yale University Press.

Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2008. "Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16(4 SPEC. ISS.):372–403.

Osman, Ibrahim H, Abdel Latef Anouze, Zahir Irani, Baydaa Al-Ayoubi, Habin Lee, Asım Balcı, Tunç D Medeni and Vishanth Weerakkody. 2014. "COBRA framework to evaluate e-government services: A citizen-centric perspective." *Government Information Quarterly* 31(2):243–256.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand. 2014. "Structural topic models for open-ended survey responses." *American Journal of Political Science* 58(4):1064–1082.

Roberts, Margaret E, Brandon M Stewart and Edoardo M. Airoldi. 2016. "A model of text for experimentation in the social sciences." *Journal of the American Statistical Association* pp. 1–49.
  URL: *http://www.tandfonline.com/doi/full/10.1080/01621459.2016.1141684*

Rosen-Zvi, M., T. Griffiths, M. Steyvers and P. Smyth. 2004. "The author-topic model for authors and documents." *Proceedings of the 20th conference on Uncertainty in artificial intelligence*
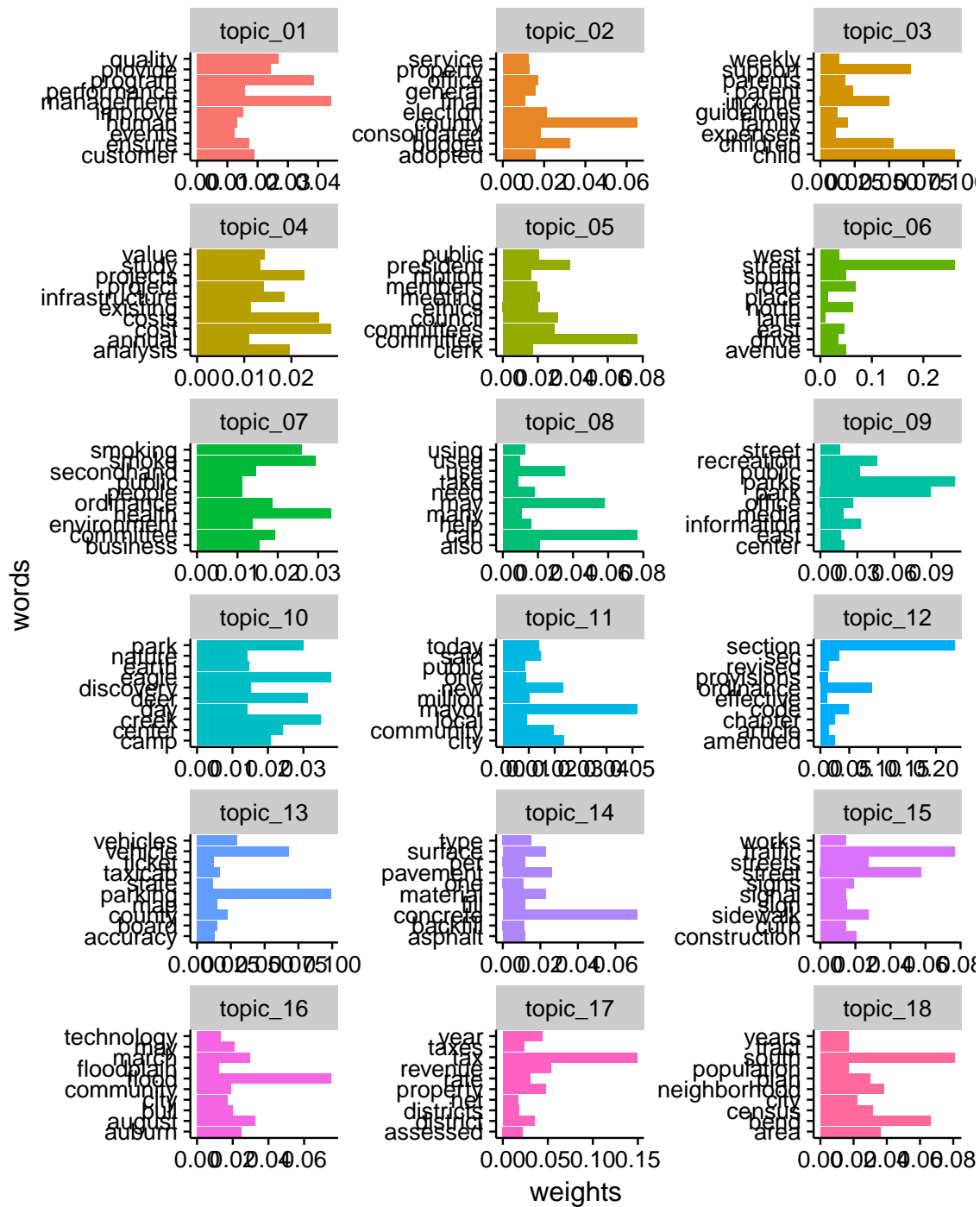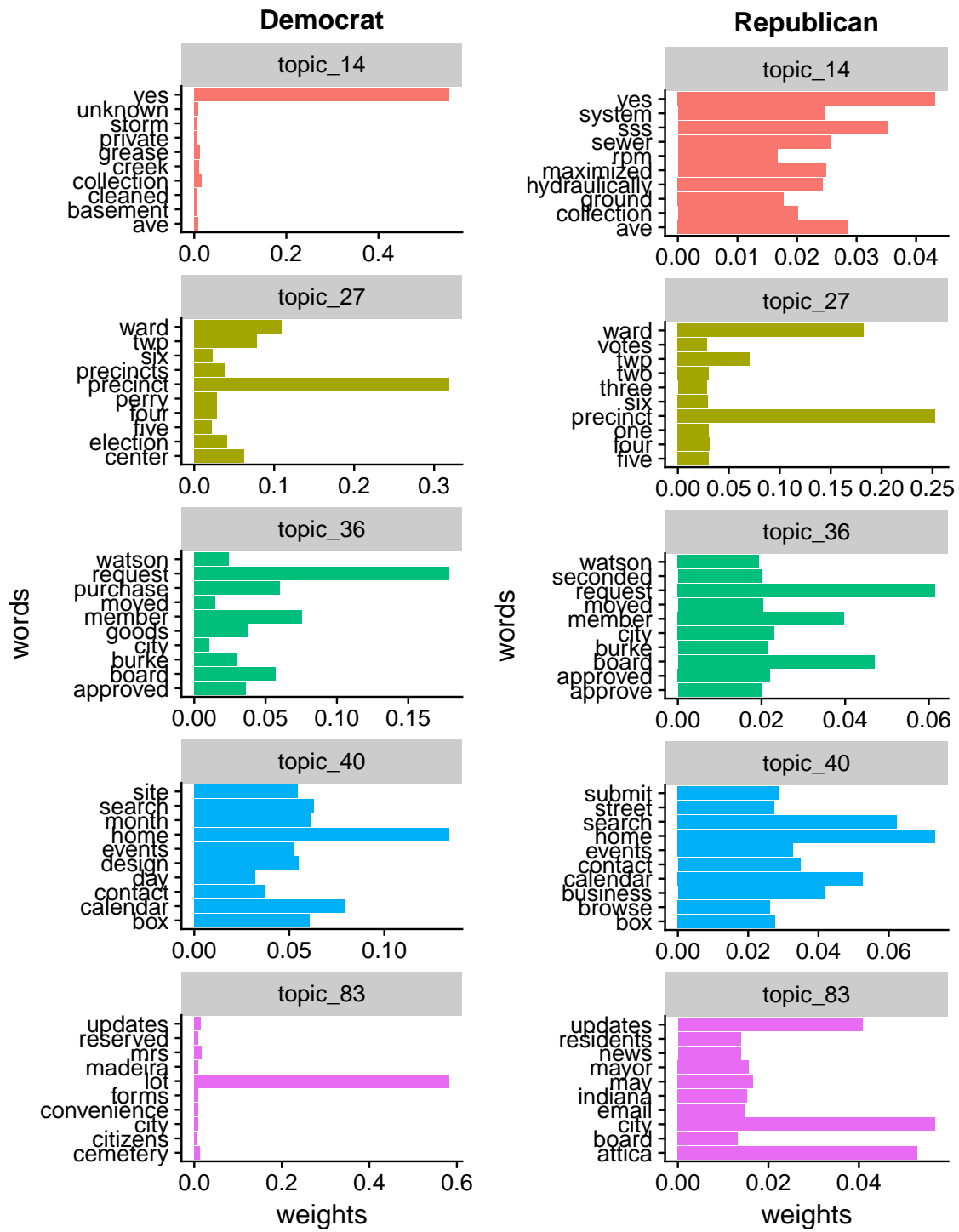
pp. 487–494.
URL: *http://portal.acm.org/citation.cfm?id=1036902*

Wallach, Hanna M, David Mimno and Andrew Mccallum. 2009. "Rethinking LDA: Why Priors Matter." *Advances in neural information processing systems* .

Wang, Lili, Stuart Bretschneider and Jon Gant. 2005. Evaluating web-based e-government services with a citizen-centric approach. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*. Ieee pp. 129b–129b.
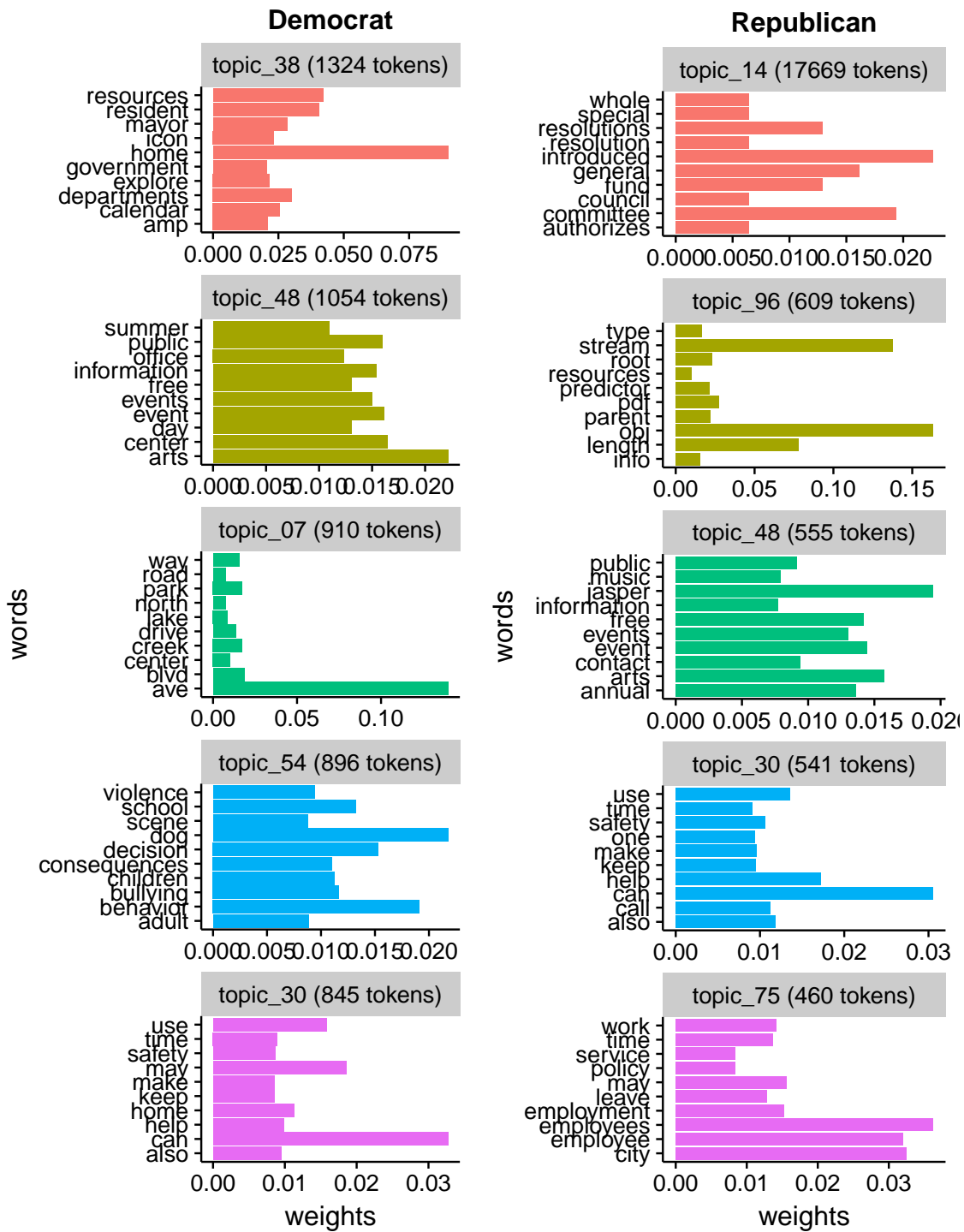
Word–topic probabilities – wget

Democrat

topic_38 (1324 tokens)

topic_48 (1054 tokens)

topic_07 (910 tokens)

topic_54 (896 tokens)

topic_30 (845 tokens)

Republican

topic_14 (17669 tokens)

topic_96 (609 tokens)

topic_48 (555 tokens)

topic_30 (541 tokens)

topic_75 (460 tokens)

words

weights

Figure 3: Densities of topic weights for documents in Republican and Democratic cities.
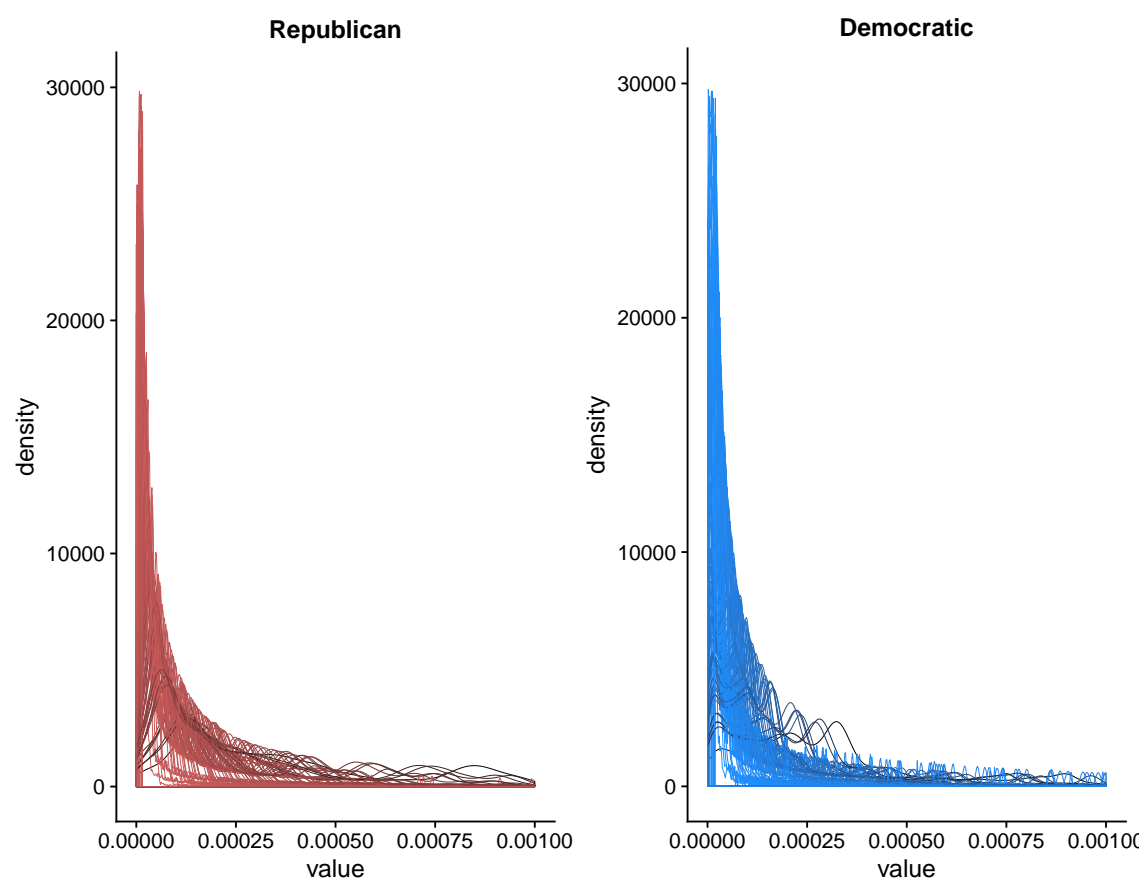
Figure 4: Densities of topic weights for documents in Republican and Democratic cities.
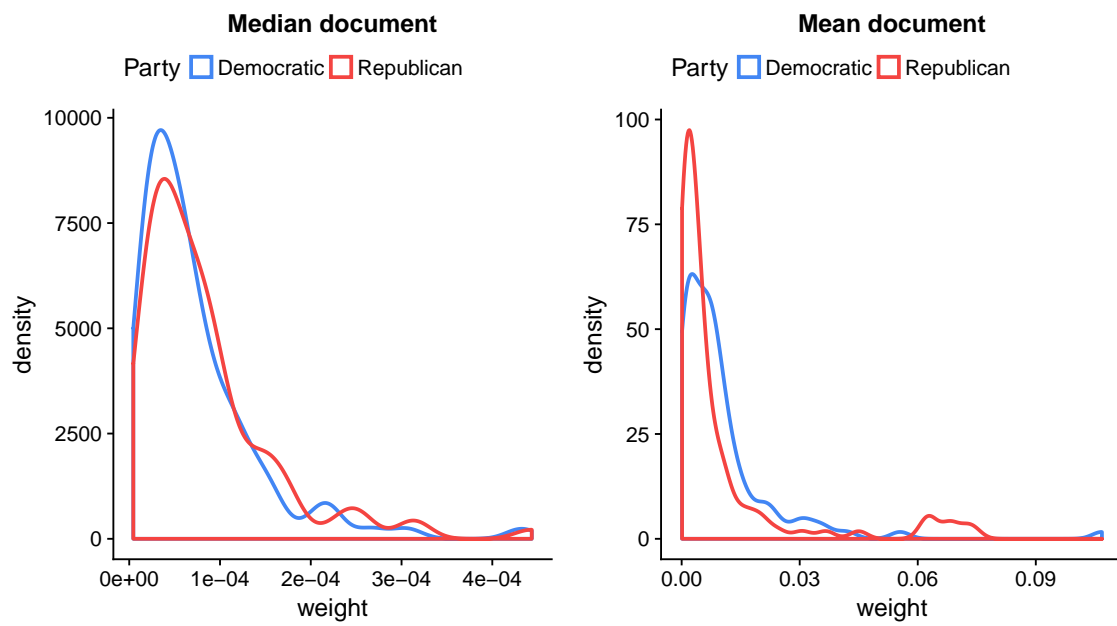
Figure 5: Word-topic probabilities for topics with big partisan differences
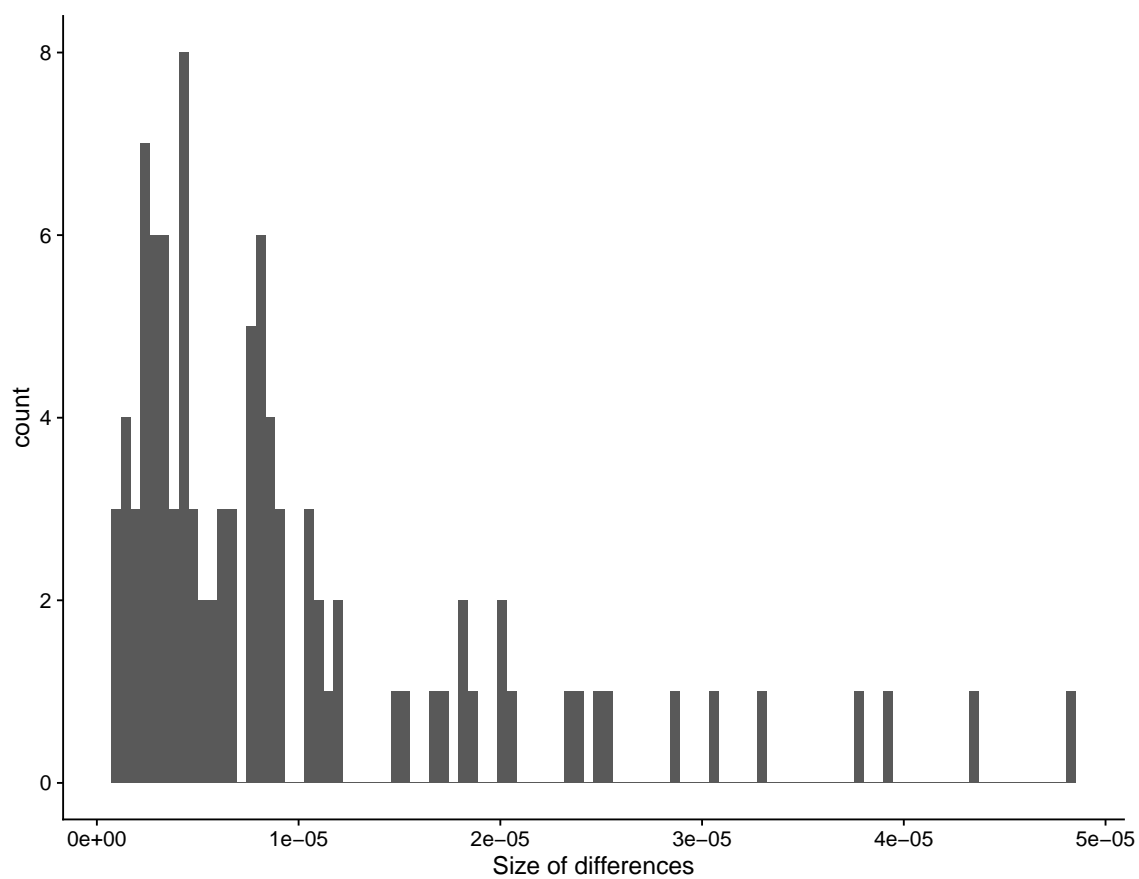
Figure 6: Word-topic probabilities for topics with big partisan differences

28
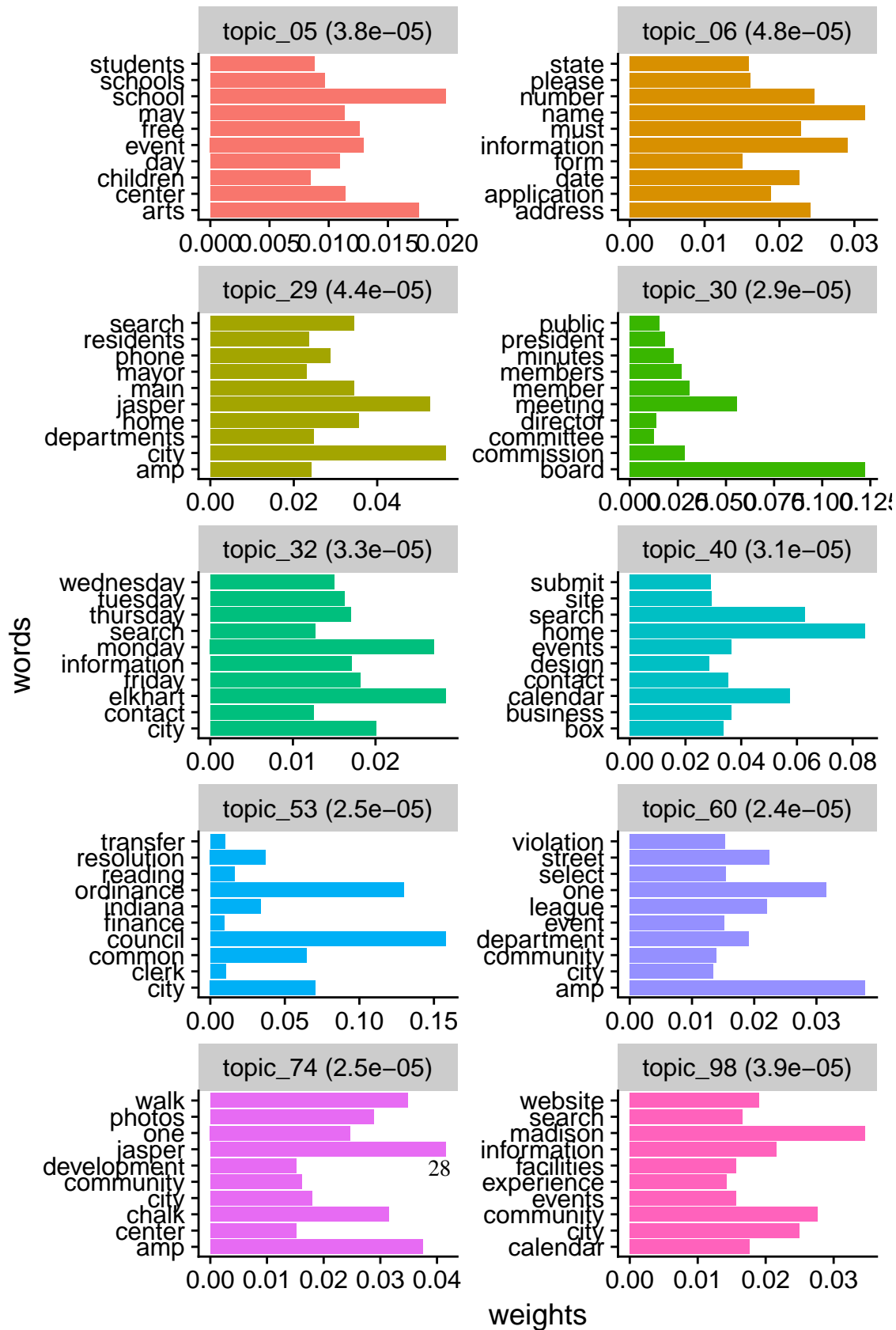
Figure 7: Word-topic probabilities for topics with big partisan differences
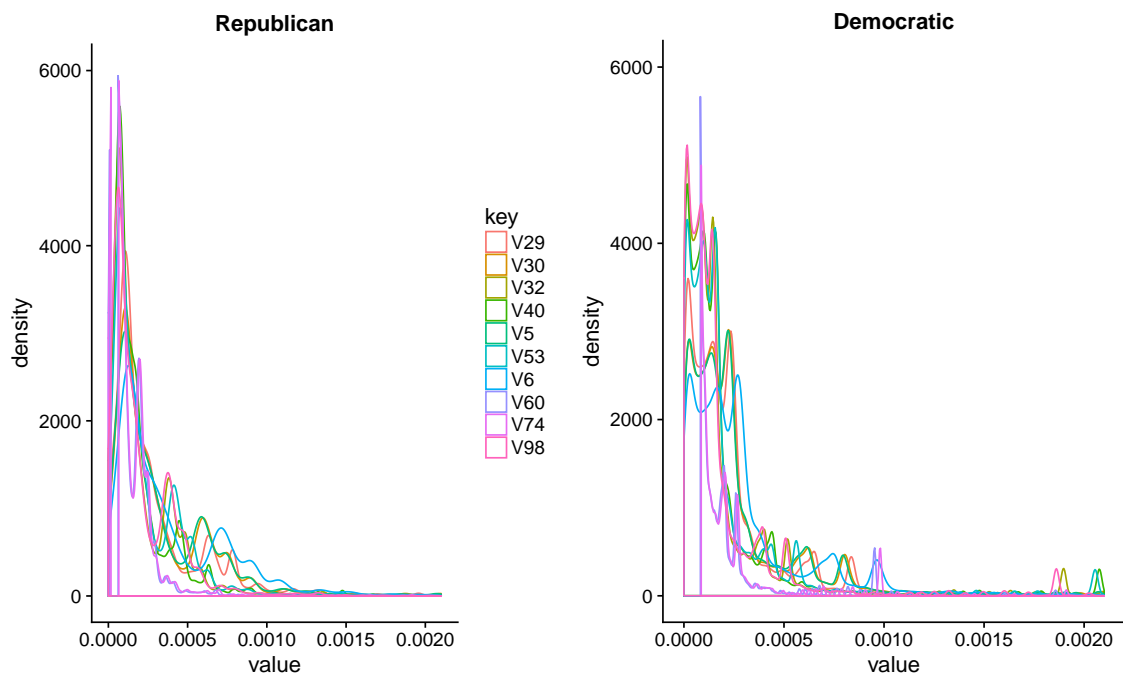
Figure 8: Word-topic probabilities for topics with big partisan differences

Democratic

Republican

30
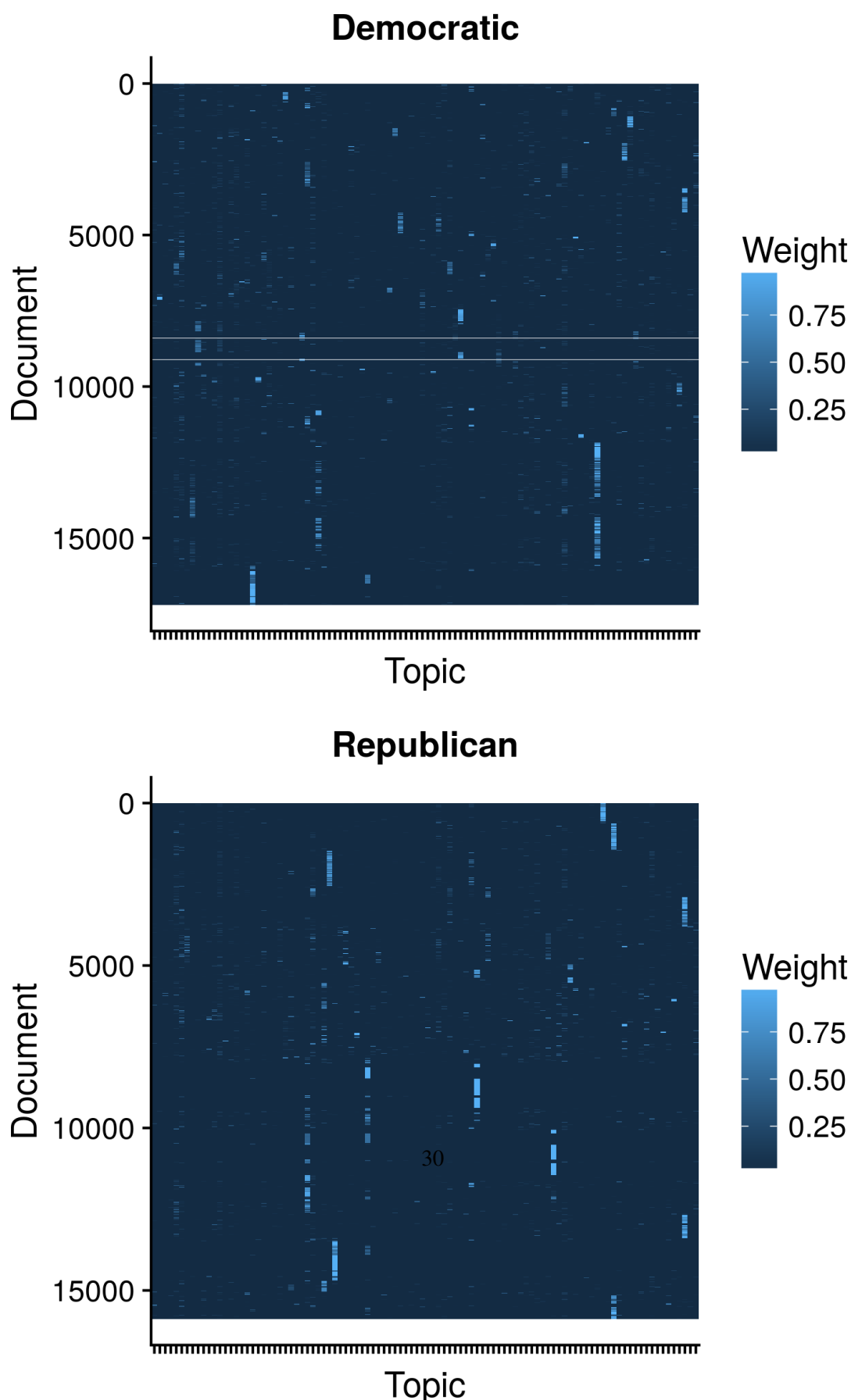
Figure 9: Topic coherence, varying the number of topics. The red line represents the mean topic coherence for each number of topics.
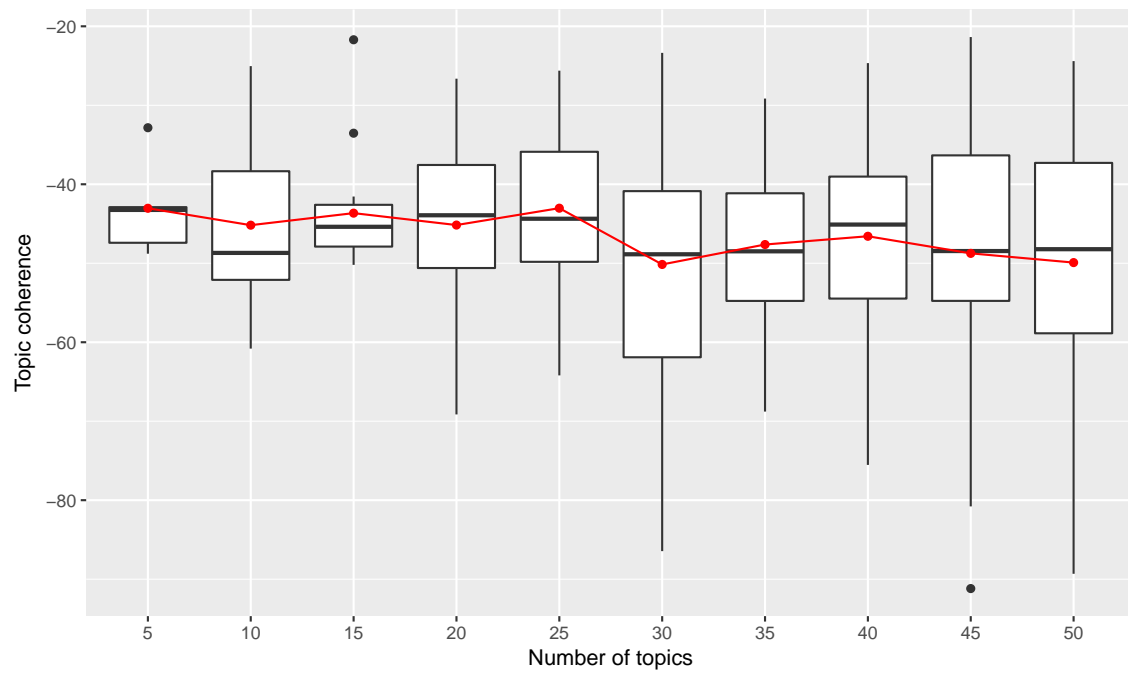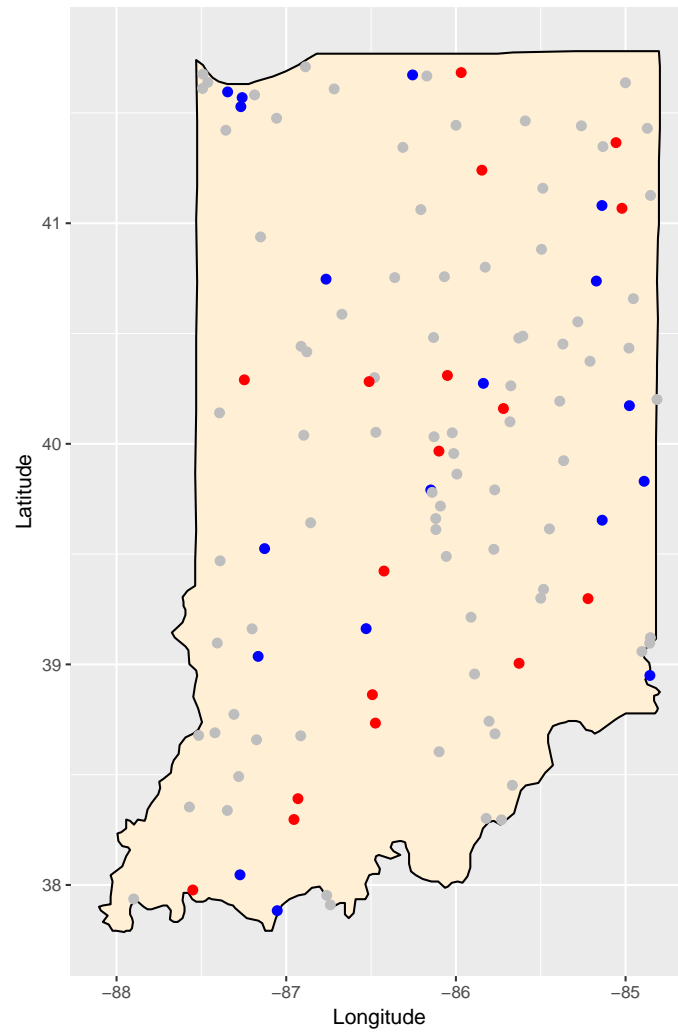
Figure 10: Cities in the corpus, by partisanship of mayor.

Figure 11: The most partisan topics. The number in parenthesis indicates how many times more the topic appears on average (measured through the number of words of the topic throughout the documents) in the respective party's corpus (indicated by the color).
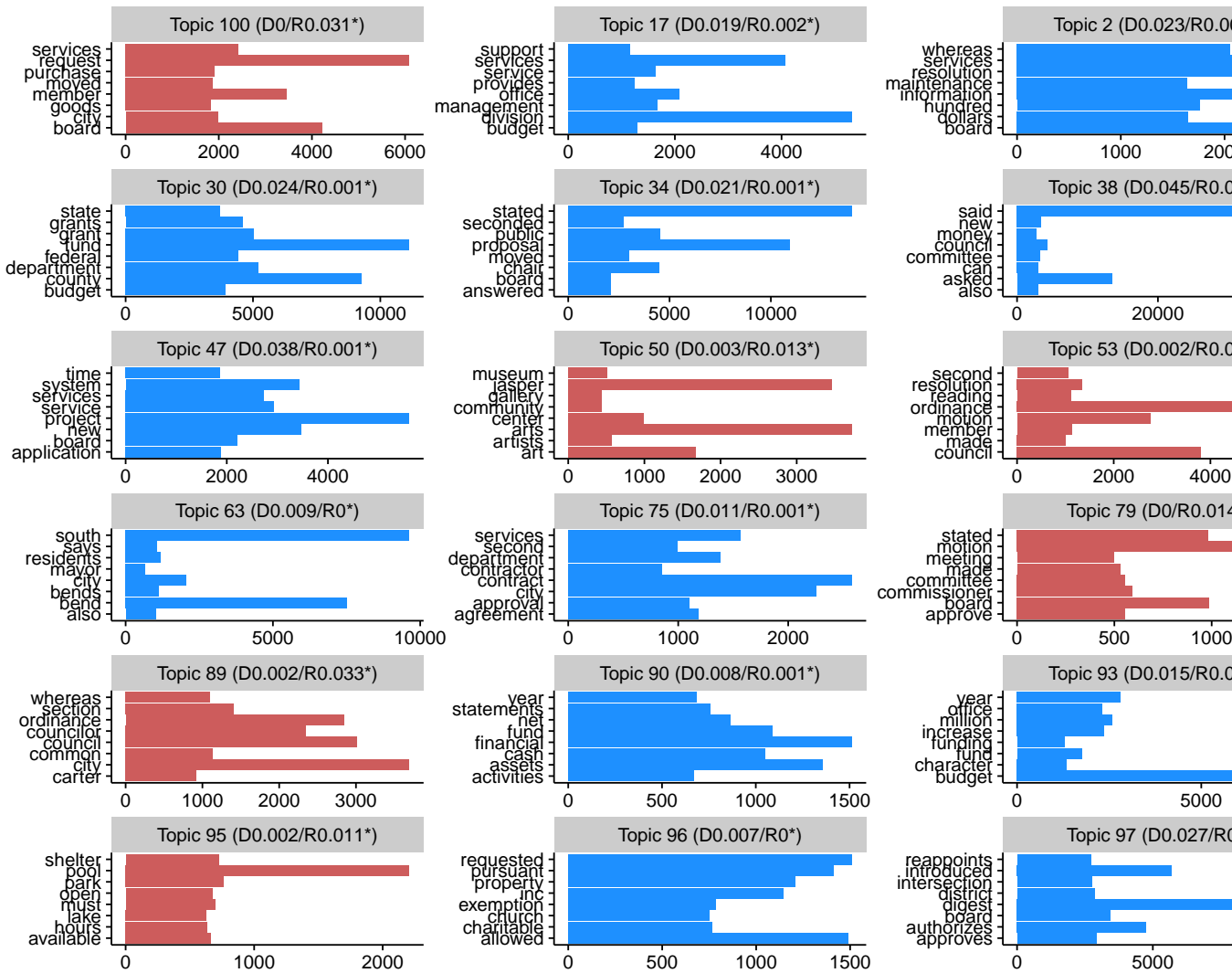
Figure 12: The optimal number of topics for our corpus. The measures of Griffths 2004 and Cao 2009 indicate the best number of topics at their minimum, whereas the measure of Arun 2010 points to the best value at its maximum.
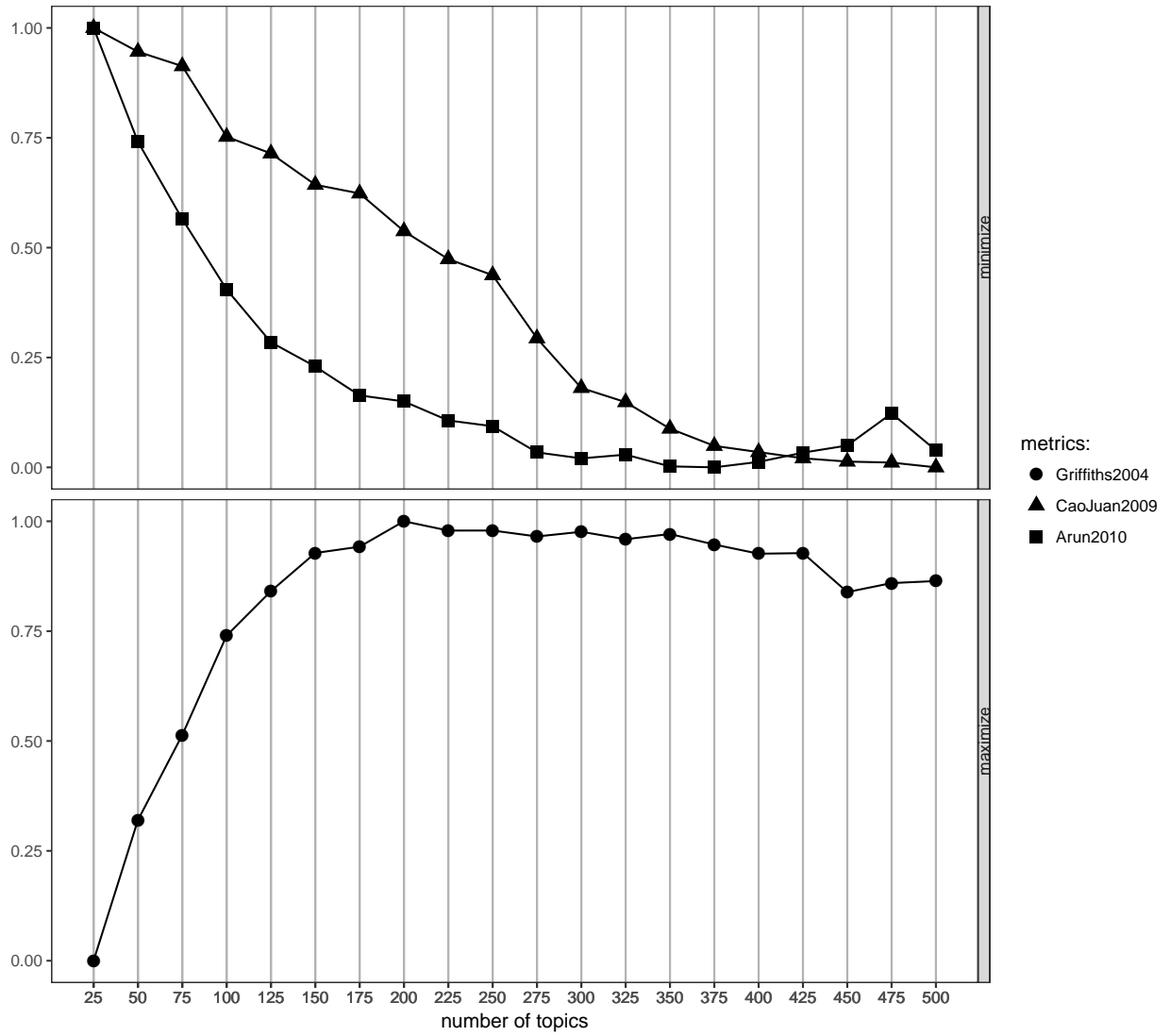
Figure 13: Results from a structural topic model, displayed as the p-values for each variable for each topic. This would normally be somewhat nonsensical, but here it illustrates why the model does not work.