The Effect of Elections and Changed Governments on Discussion in the Australian Federal Parliament (1901–2017) *

Monica Alexander *University of Toronto*
Rohan Alexander *Australian National University*

We systematically analyse how parliamentary discussion changes in response to elections and changed governments in Australian history. We first create a dataset of what was said in the Australian Federal Parliament from 1901 through to 2017 based on available public records. To reduce the dimensionality of this dataset we use a correlated topic model, and then analyse the effect of various events using a Bayesian hierarchical Dirichlet model. We find that: changes in government tend be associated with topic changes even when the party in power does not change; elections that do not result in a change in government are rarely associated with topic changes; economic events, such as financial crises, have less significant effects than other events such as terrorist attacks; and the effect of events is much more pronounced in the second half of our sample, and especially in the past two decades. Our findings have implications for how we think about the longer-term trajectory of government policymaking as the media and political cycles becomes increasingly focused on short-term events.

Keywords: text, politics, Australia, unsupervised machine learning, Bayesian model.

1 Introduction

Government policy is partly driven by parliamentary discussion. Conversely, that same discussion can indicate a government's priorities. But major events – be they expected, such as an election, and unexpected, such as a recession or terrorist attack – can affect the course of parliamentary discussion. For instance, think of how a new government often goes to some trouble to seem different to the one they replace, or how events such as the 9/11 attacks altered government priorities.

In this paper we examine text records of what was said in the Australian Federal Parliament. We create a dataset based on all sitting days from 1901 through to 2017. We have records for 7,869 days in the House of Representatives and 6,682 days in the Senate. Our dataset consists of counts of every word that occurs at least thrice, after we convert all words to lower-case, remove numbers and punctuation, join commonly co-located words, and remove stop words.

*Thank you to Chris Cochrane, Dan Simpson, Jill Sheppard, John Tang, Leslie Root, Martine Mariotti, Matt Jacob, Matthew Kerby, Tianyi Wang, Tim Hatton, and Zach Ward for their helpful suggestions; and to the UC Berkeley Demography Department for the use of their computing resources. We are grateful for the many excellent comments that we received from seminar participants at the ANU School of Politics and International Relations, the ANU Research School of Economics, and the Parliamentary Library. Comments and suggestions on the 09 November 2018 version of this paper welcome at: rohan.alexander@anu.edu.au.

We systematically analyse the text using two main statistical techniques. To reduce the dimensionality of our dataset we first use a topic model that allows correlation between topics. We then analyse those topics using a Bayesian hierarchical Dirichlet model to examine changes at various types of events. These events include: changes in government; elections; and other significant events such as an economic recession or terrorism.

We find: 1) changes in government tend be associated with topic changes even when the party in power does not change. For instance, the change from Hughes to Bruce in 1922, Menzies to Holt in 1966; Hawke to Keating in 1991; and Rudd to Gillard in 2010 are all associated with significant changes in topics despite no change in the party in power. 2) As expected, elections where the party in power also changes, such as Fisher in 1910 and 1914, Menzies in 1949, Whitlam in 1972, Fraser in 1975, Hawke in 1983, Howard in 1996, Rudd in 2007, and Abbott in 2013 are associated with topic changes, but the 1974 Whitlam, 1980 Fraser, and the 1998 and 2004 Howard re-elections, stand out as elections where the government was returned yet there was a significant change in topics. 3) Economic events, such as financial crises, have less significant effects than other events such as terrorism.

As our dataset covers 117 years we are able to see how the effect of events changes over time. We find that the events that we are interested in are rarely significant in the first half of our dataset. With a small number of exceptions, even changes of government where the party in power also changed were not associated with overly large changes in parliamentary discussion. If governments try to more thoroughly distinguish themselves from their predecessor then longer-term policy may be more difficult to enact. Similarly, there is the danger that larger-scale signature-accomplishments of a government may be neglected by their successor for solely political reasons.

Our work contributes to a growing modern quantitative social sciences literature that considers text as an input to more traditional methods, rather than requiring separate analysis. This literature sits across and draws from various historically-separate disciplines including economic history, political science, and applied statistics. We contribute to this literature in terms of both data and methods. From a data perspective we bring to bear an essentially-complete record of what was said in the Australian Federal Parliament, and we make our dataset available to other researchers via the R package `hansardr`. From a methods perspective, our analysis model has several advantages over existing methods. These include: allowing the events to have more-complicated auto-correlated functional forms; implementing pooling across groups of similar documents; and identifying outlying topic distributions without the need to pre-specify the event of interest. As the digitisation of historical sources continues and computational power becomes cheaper, we expect interest in this approach to only increase.

2 Data

Following the example of the UK a daily text record called Hansard of what was said in the Australian Federal Parliament has been made available since it was established

in 1901.[1] Earlier work on the influence of parliaments, such as Van Zanden, Buringh and Bosker (2012), often examined broader activity measures such as counts of sitting days. This allowed for long time-frames and wide comparisons. But analysing Hansard records and their equivalents directly is increasingly viable as new methods and reduced computational costs make it easier.

The recent digitisation of the Canadian Hansard, Beelen et al. (2017), allowed Rheault and Cochrane (2018) to examine ideology and party polarisation in Britain and Canada. In the UK, Duthie, Budzynska and Reed (2016) analysed Hansard records to examine which politicians made supportive or aggressive statements toward other politicians between 1979 and 1990 and Peterson and Spirling (2018) examined polarisation. And as digitisation methods improve older UK records can be analysed, for instance Dimitruk (2018) considers the effect of estate bills on prorogations in seventeenth century England. In New Zealand, Curran et al. (2017) modelled the topics discussed between 2003 and 2016, and Graham (2016) examined unparliamentary language between 1890 and 1950. And in the US, Gentzkow, Shapiro and Taddy (2018) examined congressional speech records from 1873 to 2016 to find that partisanship has risen in the past few decades.

Parts of Australian Hansard records have been analysed for various purposes. For instance, Rasiah (2010) examined Hansard records for the Australian House of Representatives to examine whether politicians attempted to evade questions about Iraq during February and March 2003. Gans and Leigh (2012) examined Australian Hansard records to associate mentions by politicians of certain public intellectuals with neutral or positive sentiment. And Salisbury (2011) examines unparliamentary behaviour. Fraussen, Graham and Halpin (forthcoming) examined Australian Hansard records to assess the prominence of interest groups. The closest research to ours that we have found is Boulus (2013) who examines parliamentary debate in Australia for the period 1946 to 2012.

The Australian Federal Parliament makes daily Hansard records available online as PDFs and these are considered the official release. Additionally, XML records are available in some cases.[2] We provide an example of a Hansard PDF page in Appendix A.1. There are 14,551 days of publicly available Hansard records across both chambers of the Australian Federal Parliament that we have PDFs for and further summary statistics for this are provided in Appendix A.2. Our data cleaning process indicates concerns with a small number of PDFs and these are detailed in Appendix A.3.

We use the official PDF release and the formatting of the Hansard records changes over time. We use scripts written in R (R Core Team (2018)) to convert the PDFs into daily text records. An example of the workflow and some reduced-detail scripts are provided

[1] While Hansard is not necessarily verbatim, it is considered close enough for text-as-data purposes. For instance, Mollin (2008) found that in the case of the UK Hansard the differences would only affect specialised linguistic analysis. Edwards (2016) examined Australia, New Zealand and the UK, and found that changes were usually made by those responsible for creating the Hansard record, instead of the parliamentarians. As those who create Hansard are tasked with creating an accurate record of proceedings, this suggests the records should be fit for the purpose of our analysis.

[2] Tim Sherratt makes these XML records available as a single download and also presents them in a website (http://historichansard.net/) that can be used to explore Commonwealth Hansard records from 1901 to 1980. Commonwealth XML records from 1998 to 2014 are available from Andrew Turpin's website, and from 2006 through to today from Open Australia's website. The records can also be downloaded from the Australian Hansard website.

in Appendix A.4. Some error is introduced at this stage because many of the records are in a two-column format that need to be separated, and the PDF parsing is not always accurate especially for older records. An example of the latter issue is that 'the' is often parsed as 'thc'. These errors are corrected when they occur more than twice and can be identified.

The percentage of stop-words in each record is reasonably consistent over time. This suggests that there is no significant difference in the quality of the parsing over time. Details of this check are provided in Appendix A.5. We use Hansard records on a daily basis in this paper. We pre-process our text before applying a topic model. The specific steps that we take are to: remove numbers and punctuation; change the words to lower case; and concatenate multi-word names titles and phrases, such as new zealand to new_zealand. Then the sentences are de-constructed and each word considered individually.

3 Model

The goal of our modelling strategy is twofold. Firstly, we want to use topic modelling (Blei, Ng and Jordan, 2003) to summarise the Hansard text into meaningful topics that reduce the dimensionality of the text data and capture the main themes discussed in parliament over time. Secondly, we want to relate the resulting topic distributions to temporal trends, changes, and events, such as a change in government, elections, and other external events. There are two stages in our analytical process.

We first use the Correlated Topic Model (CTM) (Blei and Lafferty, 2007) to obtain estimated topic distributions over time. We consider these topic distributions as reduced dimension inputs that can be analysed within another model. We then formulate a Bayesian hierarchical Dirichlet model to assess changes in the topic distributions in relation to events of interest.

In the following section, we briefly describe the topic modelling approach, which is to use the CTM, which is a natural extension of the Latent Dirichlet Allocation (LDA) model (Blei, Ng and Jordan, 2003), before discussing the Bayesian hierarchical Dirichlet analysis model used to investigate changes in topics. More detail on the topic modelling is available for readers that may be less familiar with it in Appendix B.

3.1 Latent Dirichlet Allocation

Although more- or less-fine levels of analysis are possible, here we are primarily interested in considering a day's topics. This means that each day's Hansard record needs to be classified by its topics. Sometimes Hansard records includes titles that make the topic clear. But not every statement has a title and the titles do not always define topics in a well-defined and consistent way, especially over longer time periods.

Other work such as Baumgartner and Jones (1993) and Dowding et al. (2010) does this by creating a standardised codebook of policy categories and sub-categories and then manually assigning text to topics as appropriate. This approach ensures the categorisation is reasonable but as it is a manual process the size of the text that can be categorised is limited. In exchange for some reduction in the reasonableness of the categorisation, the

LDA method of Blei, Ng and Jordan (2003) is able to provide consistent categorisation of the topics discussed in Hansard for large text collections.

The key assumption behind LDA is that each day's text, 'a document', in Hansard is made by speakers who decide the topics they would like to talk about in that document, and then choose words, 'terms', that are appropriate to those topics. A topic could be thought of as a collection of terms, and a document as a collection of topics, where these collections are defined by probability distributions. The topics are not specified *ex ante*; they are an outcome of the method. In this sense, this approach can be considered unsupervised machine learning. Terms are not necessarily unique to a particular topic, and a document could be about more than one topic. The goal is to have the words found in each day's Hansard group themselves to define topics. This can provide more flexibility than other approaches such as a strict word count method, but can require a larger dataset and make interpretation more difficult. More detail on how LDA works is available in Appendix B.

One weakness of the LDA method is that it considers a 'bag of words' where the order of those words does not matter (Blei, 2012). It is possible to extend the model to reduce the impact of the bag-of-words assumption and add conditionality to word order. Additionally, alternatives to the Dirichlet distribution can be used to extend the model to allow for correlation. This is the Correlated Topic Model, described in the next section.

3.2 Correlated Topic Model

As mentioned in the previous section, one of the limitations of LDA is that the model assumes that the presence of one topic is not correlated with the presence of another topic. Correlation between topics is neither modelled nor accounted for by LDA, but in reality often topics are related. For instance, in the Hansard context, we may expect topics related to the army to be more commonly found with topics related to the navy, but less commonly with topics related to banking. The goal of the CTM (Blei and Lafferty, 2007) is to account for this correlation between topics, in order to produce more realistic and stable topic distributions over time. The models are very similar, and the key difference is the underlying distributions that are drawn from.

As with LDA, the process assumed to generate the documents is the key aspect as this will be reversed to estimate the topics. The document generation process of Blei, Ng and Jordan (2003) discussed earlier, is just slightly modified. Rather than assuming that the distribution of topics in a document, θ_d, are a draw from a Dirichlet distribution, as in step 2 in LDA above, CTM assumes

$$\theta_d \sim \text{Logistic Normal}(\mu, \Sigma)$$

That is, the main difference of CTM over LDA is that it replaces the assumption of the Dirichlet distribution with a more flexible logistic multivariate Normal distribution. This distribution can incorporate a covariance structure across the topics. The remainder of the steps of the document generating process are identical to LDA.

However, the replacement of the Dirichlet distribution with the logistic multivariate Normal distribution adds a level of computational complexity to CTM. The posterior distributions of the parameters of interest ($\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}$) can no longer be obtained

using standard simulation techniques such as Gibbs Sampling. Blei and Lafferty (2007) develop a fast variational inference procedure for estimating the CTM. CTM itself has also been extended by Roberts, Stewart and Airoldi (2016) as part of their work on Structural Topic Models. The main difference is to add a covariate to μ which allows consideration of additional information.

Figure **??** illustrates a CTM output based on a five per cent sample from the Australian House of Representatives between 1901 and 2017. It shows how each day's parliamentary discussion can be apportioned to a topic and highlights how these proportions change over time.

3.3 Analysis model

The CTM output of interest is the proportion of each topic appearing in each document, that is, the θ_d. The aim of this stage of the modelling process is to analyse how the distribution of topics changes in relation to different types of events. But with 80 topics for each of the roughly 14,551 chamber-sitting-days the data are still too noisy to easily visualise changes around events.

One option for relating the topic distributions to events would be to use the Structural Topic Model (STM) of Roberts, Stewart and Airoldi (2016). The distinguishing aspect of the STM is that it considers more than just a document's content when constructing topics. For example, we may believe a document's author, or the time at which it was written, or, in the case of Hansard, the government or election period, may affect the topics within that document. The STM allows this additional information, or metadata, to affect the construction of topics, though influencing either topical prevalence or topical content. The assumption that there is some document generation process is the same as in LDA, it is just that this process now includes metadata.

More specifically, consider a matrix of covariates. X_d, where each column relates to a different covariate or metadata aspect, and each row refers to a document. Then a cell has the value of each covariate for a particular document. Similar to the CTM, the STM assumes the topics within a document θ_d are a draw from a logistic Normal distribution with mean μ. However, the STM framework assumes that the mean parameter is a linear function of covariates X_d:

$$\theta_d | X_d \gamma \Sigma \sim \text{Logistic Normal}(\mu = X_d \gamma, \Sigma)$$

Using the STM framework with covariates could theoretically allow the relationships between topics and certain types of events to be assessed. For instance, one covariate could be the government g that was in power during the time corresponding to document d. However, the STM covariate framework has several limitations in terms of our goal to assess the relationship between topics and events:

- There is no way of specifying more complicated auto-correlated functional forms of the effects of events over time. For example, we believe that the effect of an election would peak at the time of the election, then gradually decay as a function of days since election. In the STM framework, it is possible to specify a constant or linear effect of elections over time, or a spline relationship over elections, but it is

not possible to restrict the effect of a specific election over time to be monotonically decreasing.

- There is no way to implement partial pooling across groups of similar documents. The STM framework assumes that documents are independently and identically distributed, conditional on the model covariates. However, it could be expected that topic distributions within a particular government, for example, may be more- or less-likely to contain certain topics for reasons that are not reflected in the topic prevalence covariates. To account for this, we would like a covariate model that allows for the partial pooling of variance in topic distributions by group, such as sitting period.
- There is no way of identifying 'outlying' topic distributions – and therefore events that had an important effect – without pre-specifying the event of interest in the model. For example, if we think that the 9/11 attacks had an effect on parliamentary discourse, then a dummy for 9/11 would have to be included in the STM framework, but the specifics of the dummy construction affect the results. Ideally, we would like to identify important events based on different-to-expected topic distributions, after accounting for time trends, government and election effects.

To overcome these challenges, we formalise a statistical framework that allows us to systematically identify significant changes in topic distributions over time. Specifically, we use the estimated topic distributions from the CTM described in the previous section as an input into a Bayesian hierarchical Dirichlet regression framework, which relates the proportions of each topic to underlying time trends, changes in governments and elections. This set-up also allows us to identify 'outlying' topic distributions and relate these to other events.

Define θ_{dp} to be the proportion of topic of topic p on day d. Note that the $\theta_{d,1:P}$ for $p = 1, 2, \ldots P = 40$ are equal to the estimated values of θ_d from the CTM. We assume that the majority of variation in topics is across sitting periods s, where a sitting period is defined as any group of days that are less than one week apart. Using this definition, there are a total of 745 sitting periods over the period 1901 to 2017 inclusive.

The topic proportions on day d are modelled in reference to their membership of a particular sitting period s. Firstly, we assume that each distribution of topics, $\theta_{d,1:P}$ for each day is a draw from a Dirichlet distribution with mean parameter $\mu_{s[d],1:P}$:

$$\theta_{d,1:P} \sim \text{Dirichlet}(\mu_{s[d],1:P})$$

where the notation $s[d]$ refers to the sitting period s to which day d belongs. This distributional assumption accounts for the fact that on any given day, the sum of all proportions in each topic must equal 1.

The goal of the model is to relate these proportions to government g at time d, and also the days since the most recent election, e, while account for underlying time trends. The mean parameters $\mu_{s,p}$ are modelled on the log scale as

$$\log \mu_{s,p} = \alpha_{g[s],p} + \cdot\alpha_{e[s],d,p} + \sum_{k=1}^{K} \beta_{p,k} \cdot x_{s,k} + \delta_{s,p}$$

where:

- $\alpha_{g[s],p}$ is the mean effect for government g (which covers sitting period s) and topic p;
- $\alpha_{e[s],d,p}$ is the effect of election e (which occurs in sitting period s) for topic p on day d since the election;
- $\sum_{k=1}^{K} \beta_{p,k} \cdot x_{s,k}$ is the underlying time trend, modelled using splines: $x_{s,k}$ is the kth basis spline in sitting period s and $\beta_{p,k}$ is a coefficient on the kth basis spline; and
- $\delta_{s,p}$ is a structured random, or levels, effect for each sitting period and topic.

The government term $\alpha_{g[s],p}$ assumes there is some underlying mean effect of each government on the topic distribution. We place uninformative priors on each of these parameters:

$$\alpha_{g[s],p} \sim \text{Normal}(0, 100)$$

The election term $\alpha_{e[s],d,p}$ assumes there is an initial effect of an election on the topic distribution, which then decays as a function of days since election, d. In particular, we model this as an AR(1) in d:

$$\alpha_{e[s],d,p} = \rho_{e[s],p} \cdot \alpha_{e[s],d-1,p}$$

One advantage of our model over using the STM is that we can restrict the effect of an election to be monotonically decreasing. This allows us to identify differences between government and election effects even when there is a one-term government. The value of the initial effect, $\alpha_{e[s],0,p}$, and the AR(1) term, ρ, both have non-informative priors:

$$\alpha_{e[s],0,p} \sim \text{Normal}(0, 100)$$

$$\rho_{e[s],p} \sim \text{Uniform}(0, 1)$$

We model the underlying time trend in topics using splines regression. The intuition behind this term is to capture the underlying non-linear trend in topic distributions over time, which is caused by large-scale structural changes in the economy, and Australian society and culture. The $x_{s,k}$ for $k = 1, 2, \ldots, K$ are the value of cubic basis splines for sitting period s at knot point k. We place knot points every five sitting periods as this is the average length of time for a government to sit. Non-informative priors are placed on the splines coefficients:

$$\beta_{p,k} \sim \text{Normal}(0, 100)$$

Finally, the sitting period-specific random effect $\delta_{s,p}$ allows for the topic distributions in some sitting periods to be different than expected based on government and election effects. This allows us to identify large deviations away from the expected distribution, thus helping to identify the effect of other, non-government and non-election events. In addition, this set up also partially pools effects across sitting periods. The $\delta_{s,p}$ values are modelled as:

$$\delta_{s,p} \sim \text{Normal}(0, \sigma^2_{e[s],p})$$

The variance parameters $\sigma^2_{e[s],p}$ give an indication of the how the variation in topics is changing over election periods. If the estimates of the variance are larger, then there is more variation in the topics discussed within an election period. Non-informative priors are placed on the variance parameters:

$$\sigma_{e[s],p} \sim \text{Uniform}(0, 3)$$

We run the model in JAGS using the `rjags` package of Plummer (2018).

4 Results

Firstly, we describe the results of the CTM approach, which defined 40 unique topics over the period 1901-2017. We then describe the results of the Bayesian analysis model, which identified governments, elections and other events that were associated with a change in the topics discussed.

4.1 Topic modelling

We applied the CTM approach discussed in Section 3.2 on the processed Hansard text database outlined in Section 2. The main output of interest are the types of topics identified by the model, and the prevalence of each topic on of each day of parliamentary discussion.

The remainder of the results refer to a topic model that had 80 distinct topics defined. The choice of 80 topics was made as a trade-off between the standard diagnostic tests that suggested a larger number of topics would be more appropriate, and the need for the topics to be tractable for our analysis model and understandable for us. The diagnostic tests are detailed in Appendix B.4.

Table 1 lists the top ten words associated with each of the 80 topics. The topics cover areas such as budgets, transport and infrastructure, war and conflict, health, education, agriculture, and trade. Note that some topics seem to somewhat overlap with their content: for example, topics 12, 17, 22 and 23 all relate to war and conflict. **[UPDATE THAT.]**

Table 1: Top words associated with each topic

Topic	Terms
1	women, rights, marriage, human, discrimination, law, equal, community, society, support
2	death, compensation, injury, estate, abolition, accident, injured, deaths, loss, died
3	constitution, parliament, power, powers, constitutional, referendum, convention, representatives, proposal, section
4	defence, forces, personnel, army, military, defence_force, equipment, base, aircraft, air
5	party, communist, matter, time, mckenna, communists, organization, country, position, henty
6	vietnam, countries, south, china, united_states, world, aid, asia, country, foreign
7	petition, petitioners, citizens, pray, parliament, assembled, representatives, duty, bound, undersigned
8	sugar, industry, bounty, queensland, growers, production, fruit, cotton, ton, paid
9	na, senate, president, question, greens, time, committee, support, australians, country
10	service, public, board, officers, officer, department, salary, commissioner, appointment, salaries
11	senators, ill, time, gardiner, measure, western_australia, collings, position, read, leader
12	television, broadcasting, service, stations, radio, post, services, commercial, abc, telephone
13	senate, senators, chamber, representatives, representing, business, party, week, position, public
14	workers, employees, relations, industrial, employers, workplace, employment, employer, union, business

Table 1: Top words associated with each topic *(continued)*

Topic	Terms
15	president, sympathy, public, word, world, personal, regret, presiding, standing, bias
16	commission, report, royal, royal_commission, inquiry, evidence, commissioner, commissions, body, appointed
17	tax, income, taxation, sales, treasurer, per_cent, pay, revenue, rate, taxes
18	fl, senate, question, greens, time, carbon, president, support, change, move
19	industrial, union, arbitration, workers, unions, trade, court, industry, conciliation, employers
20	matter, labor_party, question, debate, situation, time, organisation, lo, cavanagh, greenwood
21	department, matter, time, question, regard, money, expenditure, connection, business, information
22	per_cent, ad, val, subitem, item, omitting, inserting, exceeding, duty, intermediate
23	court, law, high_court, justice, federal, courts, attorneygeneral, legal, judge, tribunal
24	debate, time, labor_party, issue, deputy, political, question, matter, country, process
25	life, superannuation, fund, insurance, scheme, funds, national, contributions, retirement, age
26	security, iraq, support, detention, international, intelligence, time, world, australias, terrorism
27	education, schools, students, school, university, universities, training, student, funding, children
28	war, production, country, matter, control, governments, industry, time, prices, services
29	health, medical, hospital, private, insurance, medicare, hospitals, scheme, services, public
30	defence, military, naval, training, navy, forces, officers, time, force, service
31	matter, information, letter, department, evidence, document, documents, report, statement, office
32	british, great_britain, germany, empire, trade, country, canada, new_zealand, imperial, conference
33	committee, report, parliament, committees, public, parliamentary, recommendations, time, joint, inquiry
34	immigration, country, migration, migrants, citizenship, policy, immigrants, citizens, english, countries
35	question, time, debate, standing_orders, matter, business, parliament, standing, chairman, chair
36	tasmania, queensland, western_australia, new_south_wales, south_australia, victoria, federal, south, tasmanian, premier
37	rules, december, service, hon, association, january, november, october, july, june
38	housing, building, homes, home, capital, site, houses, new_south_wales, construction, canberra
39	question, department, answer, notice, provided, services, total, staff, nil, ii
40	shipping, ships, ship, vessels, line, trade, port, vessel, ports, sea
41	amendments, subsection, section, schedule, omit, item, line, substitute, person, title
42	prime_minister, party, country, leader, parliament, policy, opposite, political, time, election
43	aircraft, aviation, air, airport, airlines, transport, civil, qantas, airline, services
44	duty, per_cent, item, duties, imported, committee, revenue, industry, article, manufacturers
45	main, electorate, million, committee, community, regional, services, per_cent, time, program
46	late, parliament, loss, time, lost, public, memorial, friend, passing, regret
47	report, senate, matter, democrats, governments, leave, per_cent, aboriginal, program, button
48	roads, road, water, railway, line, transport, construction, river, country, money
49	development, time, service, national, programme, overseas, field, matter, department, country
50	world, nations, international, united_nations, countries, treaty, peace, japan, united_states, japanese
51	tariff, industry, trade, industries, board, customs, protection, country, duties, duty
52	pension, pensions, pensioners, week, social, age, benefits, service, services, repatriation
53	community, electorate, na, time, local, support, australians, day, ms, national
54	northern_territory, territory, regulations, regulation, parliament, council, governor_general, ordinance, territories, administration
55	ill, money, position, country, time, amount, financial, matter, treasurer, dr
56	environment, heritage, nsw, environmental, conservation, community, project, management, forest, council
57	energy, gas, nuclear, fuel, change, industry, emissions, power, climate, carbon
58	question, department, matter, answer, time, notice, information, report, questions, national
59	per_cent, budget, increase, economic, country, unemployment, economy, inflation, increased, time
60	care, aged, veterans, community, services, support, home, nursing, child_care, childcare
61	bank, commonwealth_bank, banking, private, credit, money, savings, treasurer, board, trading
62	research, tobacco, scientific, fishing, disease, quarantine, science, fisheries, health, fish
63	law, person, offence, criminal, police, crime, offences, evidence, penalty, attorneygeneral
64	agreement, trade, local, grants, governments, council, development, financial, national, conference
65	democrats, issue, committee, question, time, million, pmi, report, issues, process
66	war, soldiers, service, country, returned, time, ill, forces, military, soldier
67	electoral, vote, election, voting, votes, system, party, electors, elections, candidates
68	land, settlement, property, lands, country, lease, leases, pastoral, money, acres

Table 1: Top words associated with each topic *(continued)*

Topic	Terms
69	information, amendments, support, ensure, report, services, national, review, financial, provide
70	wheat, wool, growers, industry, farmers, board, prices, scheme, marketing, production
71	industry, export, meat, market, dairy, farmers, producers, levy, wine, rural
72	expenditure, loan, increase, revenue, amount, total, budget, money, financial, estimated
73	question, time, matter, desire, regard, learned, opinion, deal, position, party
74	budget, tax, billion, million, per_cent, business, economy, support, jobs, governments
75	millen, senators, question, mcgregor, de, dobson, givens, lt, clemons, time
76	company, oil, companies, industry, coal, profits, capital, business, private, mining
77	per_cent, industry, tax, policy, time, governments, economic, system, program, national
78	clause, provision, section, proposed, agreed, committee, words, provisions, person, matter
79	aboriginal, per_cent, program, governments, commission, question, assistance, funds, development, deputy
80	family, children, families, parents, income, welfare, time, poverty, allowance, parent

4.2 Analysis model

We are interested in considering the effect of various political and other events on what is talked about in the Australian Federal Parliament. As discussion in Section 3.3, the next stage of the modelling process takes the topic proportions estimated in the CTM approach, and models the association between these topic distributions and outside events.

There are several outputs of interest from this modelling stage. For example, the model provides estimates of topic prevalence by each sitting period. This nicely illustrates how the topics change over time, as the daily estimates tend to be quite variable, but using periods defined by governments or elections tend not to provide enough variability. For instance, examining Topics 12, 17, 22 and 23, which have to do with war and conflict illustrates Australia's involvement in World War I, World War II, the Korean War, the Vietnam War, the first Gulf War, and the war in Afghanistan and the second Gulf War (Figure 1).

One of the main goals of the analysis model is to see which political events are associated with changes in the prevalence of topics over time. Political events are those related to a change of government or an election. As Australia has a parliamentary system, it is possible for the government to change without an election. We define a government based on who is the prime minister, and do not distinguish between terms or cabinet composition as is sometimes done. If a person was prime minister more than once then these are considered separately.

As detailed in Section 3.3, the model estimates a mean level effect for each government, α_g and each election, α_e. We identify differences between neighbouring governments and between neighbouring elections based on calculating 95 per cent credible intervals based on posterior samples of these respective mean effects. When these do not overlap we consider that the model finds a significant difference between either the neighbouring governments or elections.

We summarise our results in terms of governments in Table 3 and in terms of elections in Table 2. These tables focus on elections and governments that were different to the ones that preceded them.

In Figures 2 and 3 we focus on certain topics to illustrate differences between governments and elections, respectively. In the graphs, the points show the estimated value of

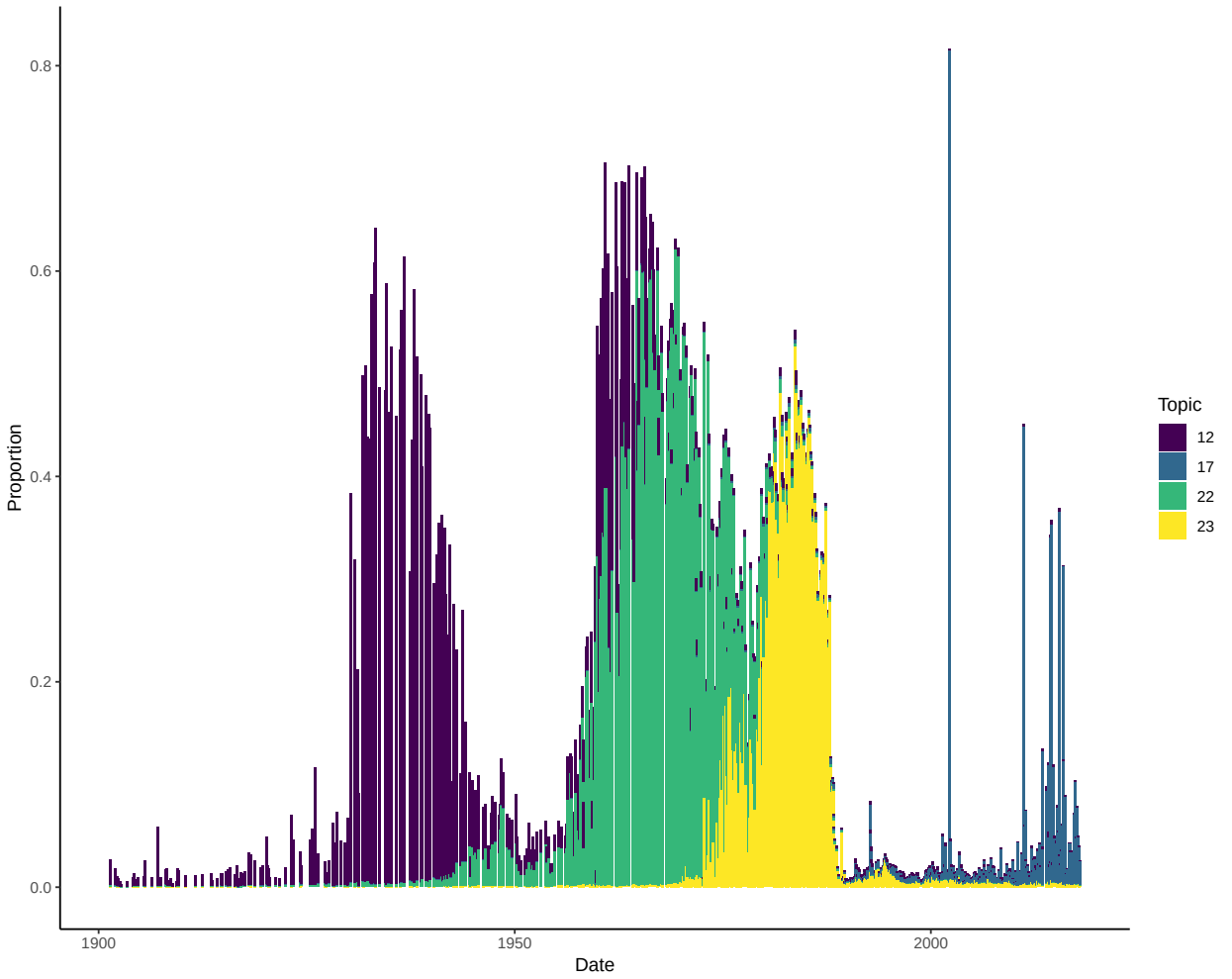


Figure 1: Model estimates of topic prevalance by sitting period for those related to war and conflict

Table 2: Significant elections, and neighours

Number	Year	Date	Total seats	Election winner	Significant
19	1949	1949-12-10	121	Non-labor	Yes
28	1972	1972-12-02	125	Labor	Yes
29	1974	1974-05-18	127	Labor	Yes
30	1975	1975-12-13	127	Non-labor	Yes
32	1980	1980-10-18	125	Non-labor	Yes
33	1983	1983-03-05	125	Labor	Yes
36	1990	1990-03-24	148	Labor	Yes
38	1996	1996-03-02	148	Non-labor	Yes
39	1998	1998-10-03	148	Non-labor	Yes
41	2004	2004-10-09	150	Non-labor	Yes
42	2007	2007-11-24	150	Labor	Yes
44	2013	2013-09-07	150	Non-labor	Yes

Table 3: Governments that were significantly different to their predecessor, and neighbours

Number	Government	Start	End	Significant
12	Bruce	1923-02-09	1929-10-22	Yes
15	Page	1939-04-07	1939-04-26	Yes
16	Menzies 1	1939-04-26	1941-08-28	Yes
21	Menzies 2	1949-12-19	1966-01-26	Yes
22	Holt	1966-01-26	1967-12-19	Yes
25	McMahon	1971-03-10	1972-12-05	Yes
26	Whitlam	1972-12-05	1975-11-11	Yes
27	Fraser	1975-11-11	1983-03-11	Yes
28	Hawke	1983-03-11	1991-12-20	Yes
30	Howard	1996-03-11	2007-12-03	Yes
31	Rudd 1	2007-12-03	2010-06-24	Yes
32	Gillard	2010-06-24	2013-06-27	Yes

α_g and α_e, respectively, for each of the topics specified. The error bars represent 95 per cent Bayesian credible intervals. The topics that we have focused on here are: 12, 17, 22, 23 which have to do with conflict, defence and security; 13 and 32 which have to do with commerce and trade; 24 and 33 which have to do with the budget, tax and spending; 29 which is to do with education and 39 which has to do with health.

Although we do not explicitly include them in the model, the non-political events that we look for when considering periods of significant change are defined by substantial changes in various economic measures, such as the onset of the Great Depression or floating the currency; events of a historical magnitude, such as entering into a war or the 9/11 attacks; or events that had a significant effect on Australian life, such as hosting the Olympics, or the Mabo decision. The full list of events that we consider are detailed in Appendix C. Note that we consider each chamber separately and future work could expand the model to better understand, and allow, for correlation between them.

We estimate sitting-period level-effects (essentially a mean for each topic by sitting period). The difference between this mean distribution and a particular day's topic distribution defines a measure that can be thought of as essentially a residual which allows us to identify outlying days. This approach means that the model generates dates that are interesting without us having to specify interesting dates.

More specifically, we define a day to be 'outlying' or 'different-to-expected' if the topic distribution on that particular day is more than three standard deviations different to the mean topic distribution for the relevant sitting period. Table 4 summarises the days where parliamentary discussion was significantly different from the rest prevailing in that week.

5 Discussion

Of the 36 different governments over this period, we find that 14 of them are significantly different to the government that preceded them. However, two of these results – the significance of the Page Government and the first Menzies Government – are likely due

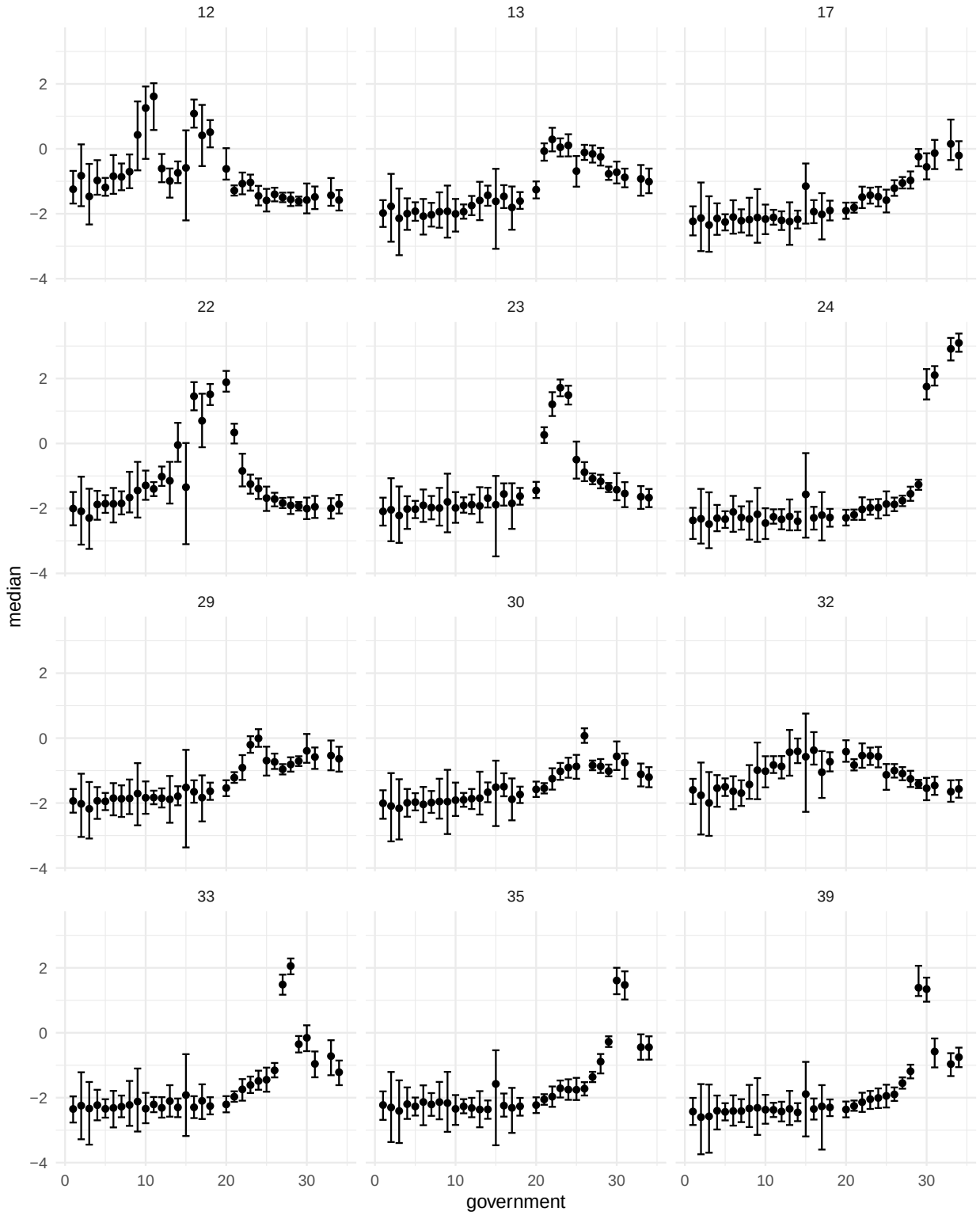


Figure 2: Government level effects on selected topics.

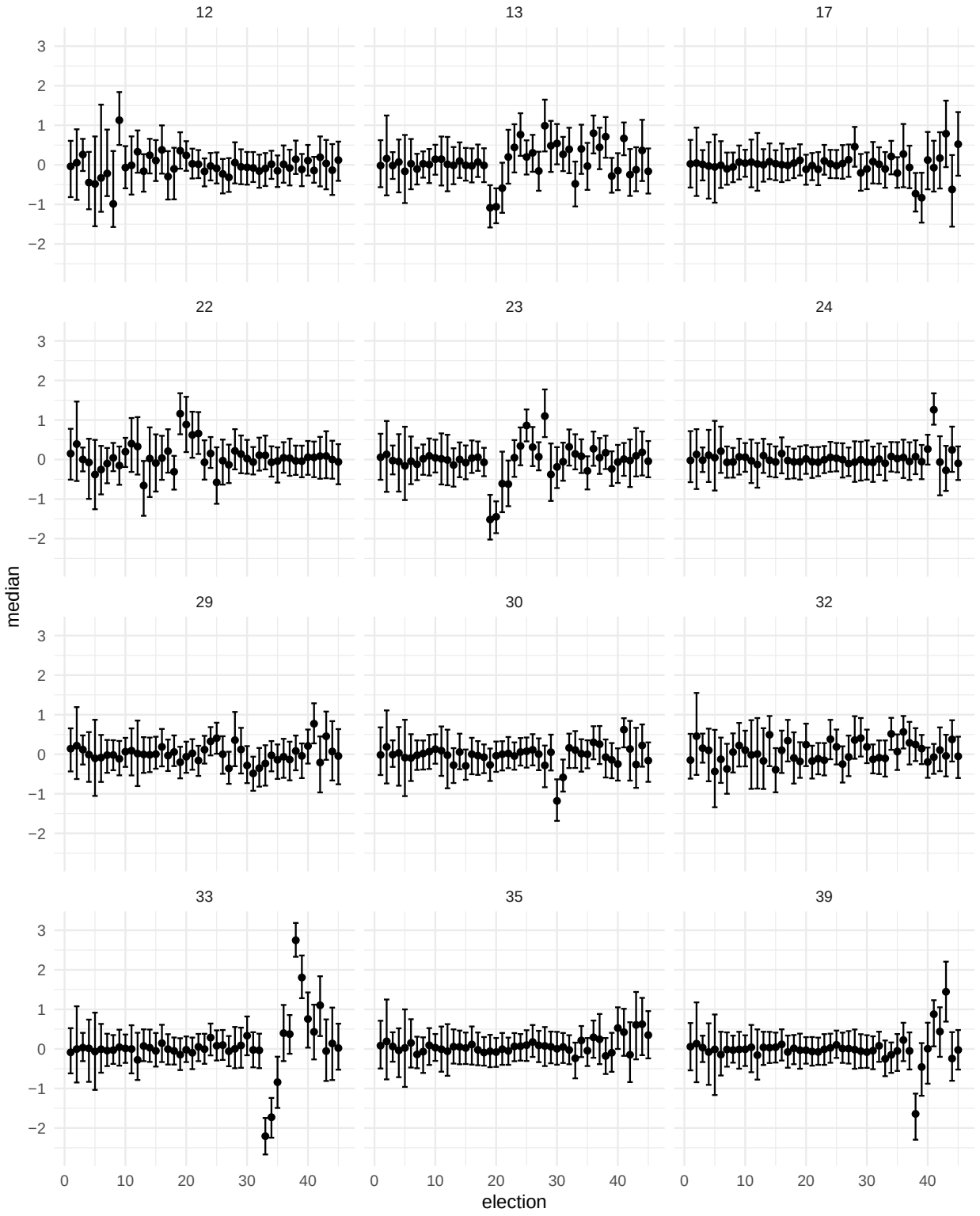


Figure 3: Elections level effects on selected topics

Table 4: Days that were significantly different to other days in their sitting period

Date	Expected
1906-08-03	No
1913-11-06	Yes
1918-11-26	Yes
1928-08-29	No
1931-06-11	No
1945-03-01	Yes
1961-09-07	No
1965-10-01	No
1970-09-18	Yes
1981-08-18	No
1982-08-17	No
2001-09-17	Yes
2002-10-14	Yes
2007-09-18	Yes
2010-02-25	No

to the short length of the Page Government and should be ignored, leaving only 12. The earliest of these are the Second and Third Fisher Governments, which were different to the Third Deakin Government and the Cook Government respectively. To a certain extent this is likely tied up with changes due to World War I.

The Second Menzies Government, starting in 1949, is the next government that is significantly different to a predecessor. The rest of the ones that are different are concentrated in the second half of our sample, with three of them being in the past twenty years. Similarly, of the 45 general elections that have been held we find that 14 of them define periods that were significantly different to their immediate predecessor. 1974, 1980, 1990, 1998, 2004, and 2007 stand out as elections where the government did not change, but the model suggests there was considerable change in the topics discussed in parliament.

The second Menzies Government was associated with a variety of changes compared with the preceding Chifley Government. The Chifley Government had governed through the end of World War II and the difficult economic times that followed. There was also a large increase in the number of seats in the House of Representatives at the 1949 election. Many new politicians entered parliament and this changed representation may also have been partly to do with the changed distribution of discussion topics. The sixteen-year length of the second Menzies Government, and better economic conditions over this time make it understandable why parliamentary discussion would have been different.

There were six elections within the second Menzies Government. However we do not find that any of those elections was associated with a significant change in the topics discussed. In this sense it was a period of consistency, especially when compared with shorter-term Whitlam Government, or the longer-term Howard Government.

The Menzies Government was succeeded by the Holt Government in January 1966. This is an example where there was a change in government without an election, as the next election only happened in November 1966. We find that the Holt Government is

significantly different to the Menzies Government. In Figures 4 we compare the topics during the final term of the Menzies Government with the topics of the Holt Government.

The Whitlam government is interesting as we find a difference in the topics after it was first elected in December 1972, compared with its second election win in May 1974. Figure 4 compares the topics that are significant in the first Whitlam term and then compares them to those in the second Whitlam government.

The Howard Government is interesting because of the significant differences between elections. For instance, each of the election periods is associated with fairly substantial differences compared with the preceding election periods, and all apart from 2001 are actually significantly different at the 95 per cent level. Figure 4 compares the topics that are significant in the different Howard terms.

To a certain extent the change after the November 2001 election is expected because of the 9/11 terrorist attacks that had only occurred two months earlier, the Bali Bombings that occurred in October 2002, and the dramatic increase in the discussion related to terrorism and conflict over these years. However the change in 1998 and 2004 is more unexpected. Although the Howard Government is the second-longest serving government and commonly thought of as a period of stability because the senior ministers were consistent as well, it might be that it is better to think of the Howard Government as a combination of three or four different periods and that the Howard Government reinvented itself over this period.

One advantage of our analysis model, compared with using the STM approach is that we can create a measure that is equivalent to testing for outliers in a model where the underlying variables were not latent. The results of this reduction in supervision are promising, but suggest specifics of our process need further refinement. For instance, our approach appropriately identifies the sitting day that first follows 11 September 2001 and 12 October 2002, which were the dates of the 9/11 Attacks and Bali Bombings respectively. It also appropriately identifies some of other key events that we were interested in. However there are many dates that we would have expected to be identified that were not, and similarly some of the dates that were identified are surprising. Further work is needed to improve this approach. For instance, our approach may not be appropriately considering step-changes.

6 Conclusion

In this paper we consider what was said in the Australian Federal Parliament between 1901 and 2017. We download and parse PDFs of Hansard to create a new dataset of text. We use a correlated topic model to group the parliamentary discussion into topics to reduce the dimensionality, and then analyse the effect of various events on the distribution of these topics using a Bayesian hierarchical Dirichlet model. In general we find that changes in government change the distribution of topics discussed in parliament, but that most elections did not. We find that significant events such as 9/11 and the Bali Bombings had a substantial effect, but that with certain exceptions, economic events did not.

By bringing a new dataset of what was said in the Australian Federal Parliament to bear for our analysis we are able to consider events over the full history of the Commonwealth of Australia. However even after cleaning the dataset remains imperfect and is

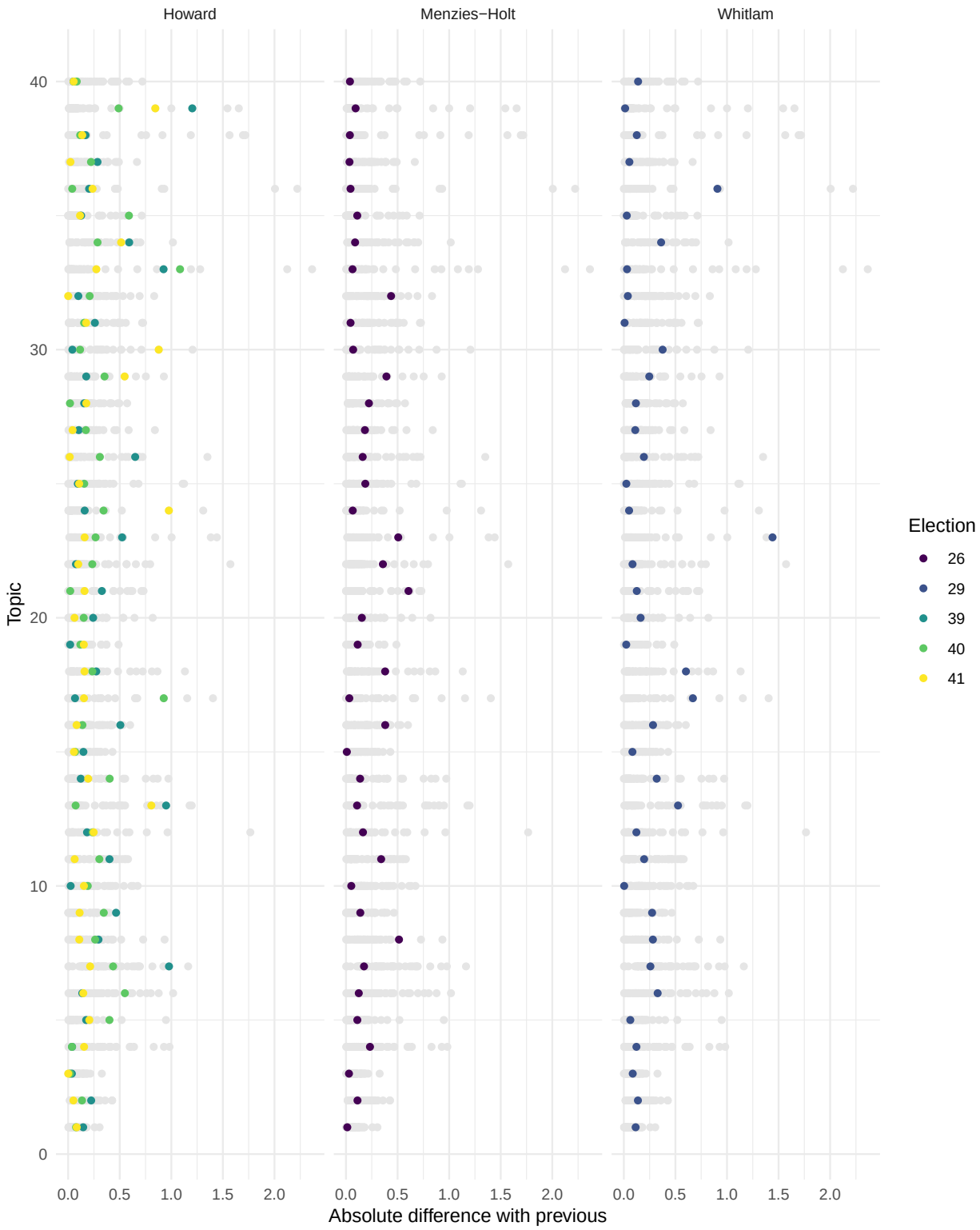


Figure 4: Differences between various elections

more fit-for-purpose than of broad applicability.

Using text as data allows larger-scale analysis that would not be viable using less-automated approaches and so it can identify associations and patterns that may otherwise be overlooked. That said, text analysis has well-known shortcomings and weaknesses and should be considered a complement to more detailed analysis such as qualitative methods and case studies. For instance, the use of topic models is problematic because of the need to interpret the topics, which at can sometimes feel like interpreting hieroglyphics. This can be difficult especially when the number of topics is large, but in a dataset of the size that we have, a large number of topics is needed. Also, although topic modelling is an unsupervised machine learning technique, the inputs require fine-tuning. For instance, selecting stop words for removal and which word to merge because of common co-location has an impact on the topics. Even after doing this there tend to be topics that are not overly meaningful, especially on their own. One way to get around using topic models is to use the words more directly, for instance word2vec and other approaches. As computational power become cheaper and more appropriate analytical methods, such as Taddy (2015) as applied in Gentzkow, Shapiro and Taddy (2018), are developed this becomes a more feasible options and future research could take that approach.

In terms of the analysis model which relates the topic distributions with events, there are several limitations to the model. Firstly, we are assuming that the effect of a particular government is constant across the whole period. In addition, we assume that the effect of elections is monotonically decreasing across days since election. In future work we will investigate different functional forms on both of these effects, and in particular try to allow for elections to have a 'lead-up' effect.

The way that we identify unusual periods could also be improved. We defined sitting periods in a constant fashion across the whole dataset time frame, but the nature of how long parliaments sit for has most likely changed. In addition, more work needs to be done on how to identify outlying events. For example, it is not clear if an important event that occurs outside a sitting period would be identified. And if an event happens in the middle of an sitting period, it may have a large effect on the overall mean, such specific days are not identified as significantly outlying.

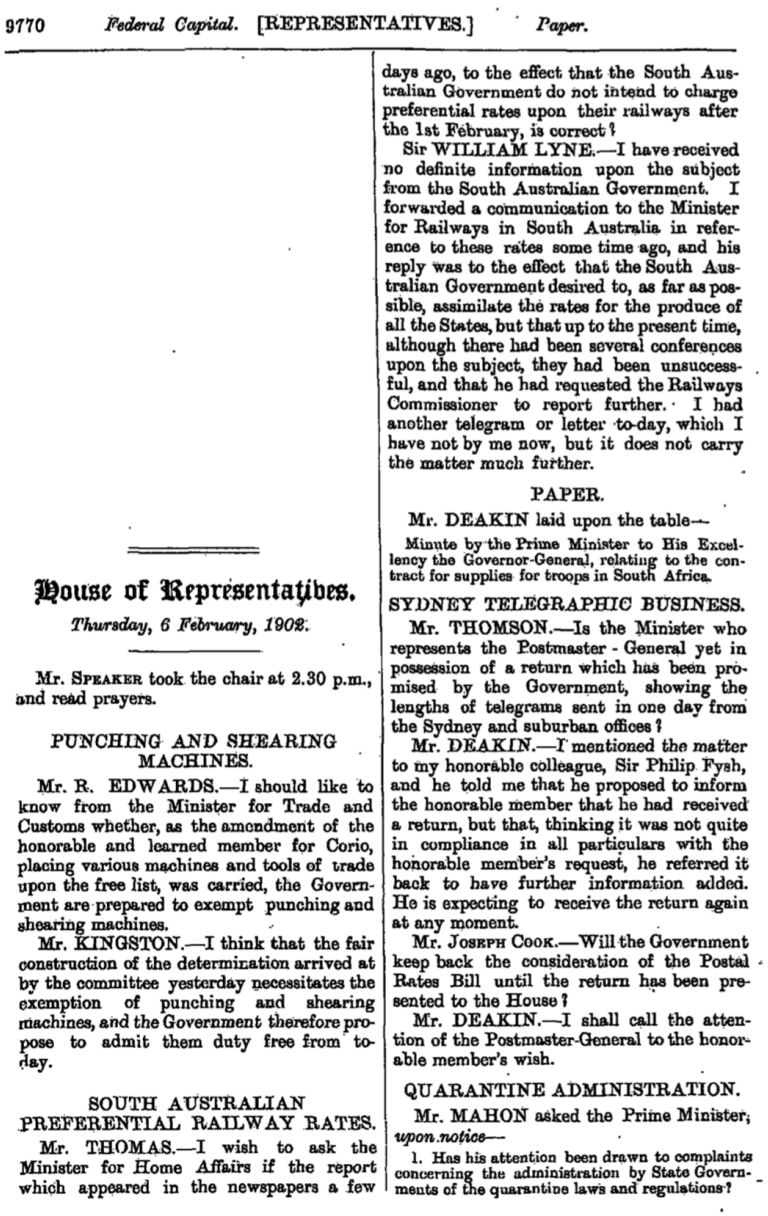
Finally, the current modelling and analysis set-up is a two-stage process: we take the output of a topic model, and use this as the input to a second model. However, this approach does not appropriately propagate uncertainty in topic distribution estimation. Future methodological work will focus on how to combine these two modelling steps by extending the STM approach into a more flexible framework.

Watching our politicians at work can sometimes be a little disheartening. It can be hard to believe that not only are those in charge shouting insults that would not be tolerated in a schoolyard, but that we voted to put them there. Nonetheless, our work suggests that reasonable debate of important topics does occur in parliament. It is easy to look back and think that we live in uniquely tumultuous times, but our analysis suggests events have always driven debate and that periods of stability may be the exception. However, we do find that since the 1980s differences between government and election periods are greater than they used to be. It is, of course, reasonable that a government should act on the mandate they receive, but it takes a long time to get policy right, and we may find progress being held back if the media and political cycles become too focused on

short-term events.

A Hansard details

A.1 Example Hansard page



9770 *Federal Capital.* [REPRESENTATIVES.] *Paper.*

House of Representatives.

Thursday, 6 February, 1902.

Mr. SPEAKER took the chair at 2.30 p.m., and read prayers.

PUNCHING AND SHEARING MACHINES.

Mr. R. EDWARDS.—I should like to know from the Minister for Trade and Customs whether, as the amendment of the honorable and learned member for Corio, placing various machines and tools of trade upon the free list, was carried, the Government are prepared to exempt punching and shearing machines.

Mr. KINGSTON.—I think that the fair construction of the determination arrived at by the committee yesterday necessitates the exemption of punching and shearing machines, and the Government therefore propose to admit them duty free from to-day.

SOUTH AUSTRALIAN PREFERENTIAL RAILWAY RATES.

Mr. THOMAS.—I wish to ask the Minister for Home Affairs if the report which appeared in the newspapers a few days ago, to the effect that the South Australian Government do not intend to charge preferential rates upon their railways after the 1st February, is correct?

Sir WILLIAM LYNE.—I have received no definite information upon the subject from the South Australian Government. I forwarded a communication to the Minister for Railways in South Australia in reference to these rates some time ago, and his reply was to the effect that the South Australian Government desired to, as far as possible, assimilate the rates for the produce of all the States, but that up to the present time, although there had been several conferences upon the subject, they had been unsuccessful, and that he had requested the Railways Commissioner to report further. I had another telegram or letter to-day, which I have not by me now, but it does not carry the matter much further.

PAPER.

Mr. DEAKIN laid upon the table—

Minute by the Prime Minister to His Excellency the Governor-General, relating to the contract for supplies for troops in South Africa.

SYDNEY TELEGRAPHIC BUSINESS.

Mr. THOMSON.—Is the Minister who represents the Postmaster - General yet in possession of a return which has been promised by the Government, showing the lengths of telegrams sent in one day from the Sydney and suburban offices?

Mr. DEAKIN.—I mentioned the matter to my honorable colleague, Sir Philip Fysh, and he told me that he proposed to inform the honorable member that he had received a return, but that, thinking it was not quite in compliance in all particulars with the honorable member's request, he referred it back to have further information added. He is expecting to receive the return again at any moment.

Mr. JOSEPH COOK.—Will the Government keep back the consideration of the Postal Rates Bill until the return has been presented to the House?

Mr. DEAKIN.—I shall call the attention of the Postmaster-General to the honorable member's wish.

QUARANTINE ADMINISTRATION.

Mr. MAHON asked the Prime Minister, *upon notice*—

1. Has his attention been drawn to complaints concerning the administration by State Governments of the quarantine laws and regulations?

Figure 5: Example Hansard page – 6 February 1902

A.2 Summary statistics

The number of sitting days in a year varies considerably. The highest in the House of Representatives was 122 days in 1904, followed by 113 days in 1901 and 1920. The year with the most sitting days in the Senate was 1902 with 93 days, followed by 1989 with 92 days, and 1986 with 86 days (Figure 6).

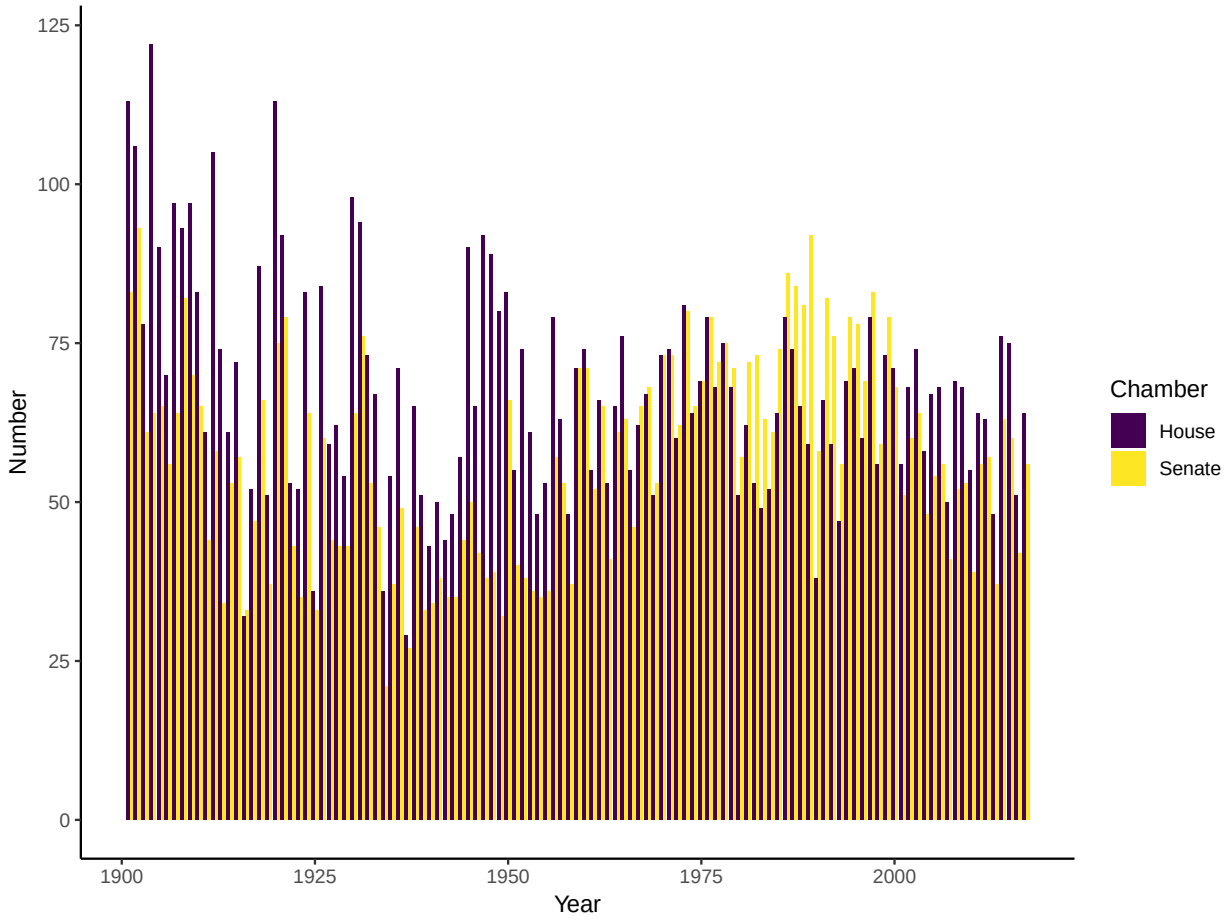


Figure 6: Number of sitting days, by year

Until the 1950s the House of Representatives tended to have more sitting days than the Senate. It was then similar, before the Senate had more days in the 1980s and 1990s. Since the 2000s the House of Representatives again has tended to have more sitting days than the Senate.

These counts of the number of sitting days are based on available PDFs. For this reason the counts may be slightly different to other counts. An example of one known issue of this type is detailed in the next section.

A.3 Compared with parliamentary website

The parliamentary website provides a summary table of the number of sitting days in each year by chamber.[3] Comparing the numbers provided in that table with number of days that we have provides an indication of how complete our dataset is.

In general the number of sitting days on the parliamentary website summary table is similar to the number of PDFs that we have although it does identify a few particularly concerning years (Figure 7).

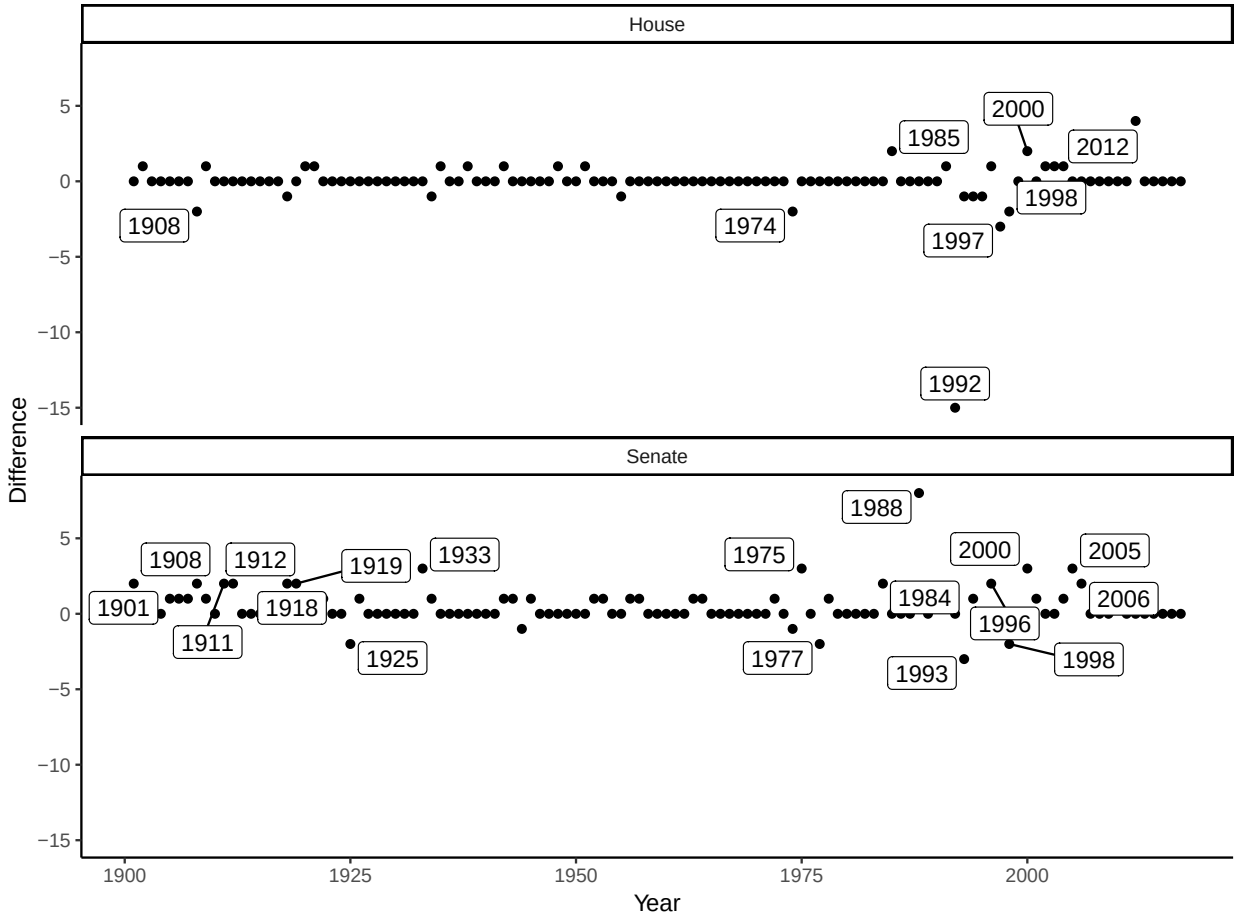


Figure 7: Example Hansard page - 6 February 1902

When the difference is positive, it means that in that year we have fewer PDFs than the parliamentary website claims. For instance, 5 could mean that the parliamentary website claimed there were 100 days, but we only had PDFs for 95 days. Similarly, when the difference is negative then we have more PDFs than the parliamentary website claims there were sitting days.

The two major years of concern are 1992 in the House of Representatives where we have 15 days more than the parliamentary website claims there were, and 1988 in the

[3]As at 5 November, the website was available at: https://www.aph.gov.au/Parliamentary_Business/Statistics/Senate_StatsNet/General/sittingdaysyear.

Senate where we have eight days fewer.

The Parliament website is missing the Hansard PDFs for the following dates in the Senate: 1988-12-21, 1988-12-20, 1988-12-19, 1988-12-16, 1988-12-15, 1988-12-14, 1988-12-13, 1988-12-12, 2000-10-12, 2000-06-19, 2004-08-09, .

There are two unaccounted for differences in 2006, one unaccounted for difference in 2001.

The Parliament website is missing the Hansard PDFs for the following dates in the House of Representatives: 1985-08-23, 1992-09-10, 1996-12-13, 2002-05-14, 2000-10-12, and 2000-06-29.

There is one unaccounted for differences in 1920, 1921, 1935, 1942, 1948, 1951, 1991, 2003, 2004, and there are two unaccounted for in 1985 and four unaccounted for in 2012

In the Senate, the PDF for 10 August 1917 may be the wrong one? It says it is for 10 Jan 1918 on the cover sheet, but there's not even an entry for 10 Jan 1918…??? Very odd….

Senate Hansard - 1918-12-18 has the wrong PDF so does 1917-08-01.

A.4 Example Hansard PDF to text record workflow

The exact scripts that we use are too long to provide here but are available on request. Instead here we provide shorterned scripts for the workflow to convert the PDFs to usable text datasets that may be a useful starting point for other researchers. The scripts are primarily based on: the `PDFtools` R package of Ooms (2018); the `tidyverse` R package of Wickham (2017); the `tm` R package of Feinerer and Hornik (2018); the `lubridate` R package of Grolemund and Wickham (2011); the `tidytext` R package of Silge and Robinson (2016); and the `stringi` R package of Gagolewski (2018). The functions of those packages are augmented by: the `furrr` R package of Vaughan and Dancho (2018); and the `tictoc` R package of Izrailev (2014). The `hunspell` R package of Ooms (2017) is used to help find spelling issues; and the `quanteda` R package of Benoit (2018) is used to compound multiword expressions.

The first step is scrape the websites for the PDFs.

```
library(purrr)
library(tidyverse)

# Create lists of URLs to scrape and file names to save the PDF as
data_to_scrape <- read_csv("csv_of_address_to_visit_and_save_names.csv")
# Pull out the two pieces
address_to_visit <- data_to_scrape$URL
save_name <- data_to_scrape$file_name
# Add a containing folder to the save_names
save_name <- paste0("pdfs/", save_name)
# Create that containing folder if necessary
dir.create("pdfs")

# The function that will visit address_to_visit and save to save_name file
visit_address_and_save_PDF <-
  function(name_of_address_to_visit,
           name_of_where_to_save) {
    # Do the actual downloading of the PDF
    download.file(name_of_address_to_visit, name_of_where_to_save)
    # Helpful so that you know progress when running it on all the records
    print(paste("Done with", name_of_where_to_save, "at", Sys.time()))
    # Space out your requests so you don't overwhelm the website
    Sys.sleep(sample(15:30, 1))
  }

# Make the function more resistent to issues
safe_visit_address_and_save_PDF <- safely(visit_address_and_save_PDF)

# Walk through the lists and get the PDFs
walk2(address_to_visit, save_name, ~ safe_visit_address_and_save_PDF(.x, .y))
```

The next step is remove the front matter from the PDFs.

```
library(furrr) # May need the devtools version devtools::install_github("DavisVaughan/fu
library(lubridate)
library(pdftools)
library(stringi)
library(tidyverse)
library(tictoc)
library(tm)
# Set up furrr to use multiple processors
plan(multiprocess)

# Create the lists of PDFs to read and file names to save text as
use_this_path_to_get_pdfs  <- "/Volumes/Hansard/pdfs/federal/hor"
use_this_path_to_save_csv_files  <- "/Volumes/Hansard/parsed/federal/hor"

# Get list of Hansard PDF filenames
file_names <-
  list.files(
    path = use_this_path_to_get_pdfs,
    pattern = "*.pdf",
    recursive = TRUE,
    full.names = TRUE
  )
file_names <- file_names %>% sample() # Randomise the order
# Modify the savenames
save_names <- file_names %>%
  str_replace(use_this_path_to_get_pdfs, "") %>%
  str_replace(".pdf", ".csv")
save_names <- paste0(use_this_path_to_save_csv_files, save_names)


#### Create the function that will be applied to the files ####
get_text_from_PDFs <-
  function(name_of_input_PDF_file,
           name_of_output_csv_file) {
    pdf_document <- pdf_text(name_of_input_PDF_file)
    # Convert to tibble so that tidyverse can be used
    pdf_document_tibble <- tibble(text = pdf_document)
    rm(pdf_document)
    # Each row is now a page of the PDF
    # Adding a column of the row numbers allows you to keep track of the page numbers
    pdf_document_tibble$pageNumbers <- 1:nrow(pdf_document_tibble)
    # Separate each line (of each page) into own row
    pdf_document_tibble <-
```

```
  separate_rows(pdf_document_tibble, text, sep = "\\n")

# Identify page headers and footers and remove them
pdf_document_tibble <- pdf_document_tibble %>%
  group_by(pageNumbers) %>%
  mutate(lineNumber = 1:n()) %>% # This gives you a line numbering for each page
  mutate(lastLine = n()) %>% # This tells you the number of the last line in each pa
  ungroup()

## Identify front matter and remove it
# Primarily identify the start of talking based on the first occurence of 'Mr SPEAKE
# It seems pretty common, but there are some misses (e.g. 1991-01-22).
# Use various backups
pdf_document_tibble <- pdf_document_tibble %>%
  mutate(
    firstSpeakerRow = str_detect(text, "SPEAKER"),
    firstPresidentRow = str_detect(text, "PRESIDENT"),
    firstJointHouseRow = str_detect(text, "JOINT HOUSE"),
    firstTookTheChairRow = str_detect(text, "took the chair"),
    firstTheChairAtRow = str_detect(text, "the chair at")
  )
pdf_document_tibble$firstSpeakerRow[pdf_document_tibble$firstSpeakerRow == FALSE] <-
pdf_document_tibble$firstPresidentRow[pdf_document_tibble$firstPresidentRow == FALSE
pdf_document_tibble$firstJointHouseRow[pdf_document_tibble$firstJointHouseRow == FAL
pdf_document_tibble$firstTookTheChairRow[pdf_document_tibble$firstTookTheChairRow ==
pdf_document_tibble$firstTheChairAtRow[pdf_document_tibble$firstTheChairAtRow == FAL

# Get the row and corresponding page and then filter to only pages from that page
row_of_first_SPEAKER <-
  pdf_document_tibble$firstSpeakerRow[pdf_document_tibble$firstSpeakerRow == TRUE] %
row_of_first_PRESIDENT <-
  pdf_document_tibble$firstPresidentRow[pdf_document_tibble$firstPresidentRow == TRU
row_of_first_JOINTHOUSE <-
  pdf_document_tibble$firstJointHouseRow[pdf_document_tibble$firstJointHouseRow == T
row_of_first_TookTheChair <-
  pdf_document_tibble$firstTookTheChairRow[pdf_document_tibble$firstTookTheChairRow
row_of_first_TheChairAt <-
  pdf_document_tibble$firstTheChairAtRow[pdf_document_tibble$firstTheChairAtRow == T

first_page_of_interest_SPEAKER <-
  pdf_document_tibble[row_of_first_SPEAKER, "pageNumbers"] %>% as.integer()
first_page_of_interest_PRESIDENT <-
  pdf_document_tibble[row_of_first_PRESIDENT, "pageNumbers"] %>% as.integer()
first_page_of_interest_JOINTHOUSE <-
```

```
      pdf_document_tibble[row_of_first_JOINTHOUSE, "pageNumbers"] %>% as.integer()
    first_page_of_interest_TookTheChair <-
      pdf_document_tibble[row_of_first_TookTheChair, "pageNumbers"] %>% as.integer()
    first_page_of_interest_TheChairAt <-
      pdf_document_tibble[row_of_first_TheChairAt, "pageNumbers"] %>% as.integer()

    first_page_of_interest_JOINTHOUSE <-
      (first_page_of_interest_JOINTHOUSE + 1)  %>% as.integer()

    filter_from_here <-
      case_when(
        !is.na(first_page_of_interest_TookTheChair) ~ first_page_of_interest_TookTheChai
        !is.na(first_page_of_interest_TheChairAt) ~ first_page_of_interest_TheChairAt,
        !is.na(first_page_of_interest_SPEAKER) ~ first_page_of_interest_SPEAKER,
        !is.na(first_page_of_interest_PRESIDENT) ~ first_page_of_interest_PRESIDENT,
        TRUE ~ first_page_of_interest_JOINTHOUSE
      )

    pdf_document_tibble <- pdf_document_tibble %>%
      filter(pageNumbers >= filter_from_here) %>%
      select(-firstSpeakerRow,
             -firstPresidentRow,
             -firstJointHouseRow,
             -firstTookTheChairRow,
             -firstTheChairAtRow)

    # Save file
    write_csv(pdf_document_tibble, name_of_output_csv_file)
    # Helpful to know where you're up to
    print(paste0("Done with ", name_of_output_csv_file, " at ", Sys.time()))
  }


#### Walk through the lists and parse the PDFs ####
# Avoid errors
safely_get_text_from_PDFs <- safely(get_text_from_PDFs)

tic("Furrr walk2 stringr")
future_walk2(file_names,
             save_names,
             ~ safely_get_text_from_PDFs(.x, .y),
             .progress = TRUE)
toc()
```

UP TO HERE

[TBD].

A.5 Stopwords over time

Insert the graph of stop words over time.

B Topic modelling example and details

B.1 Overview and example

As applied to Hansard, LDA considers each statement to be a result of a process where a politician first chooses the topics they want to speak about. After choosing the topics, the politician then chooses appropriate words to use for each of those topics. Statistically, LDA considers each document as having been generated by some probability distribution over topics. Similarly, each topic is considered a probability distribution over terms. To choose the terms used in each document, terms are picked from each topic in the appropriate proportion.

As an example, Figures 8 and 9 illustrate a smaller application with five topics, two documents, and ten terms. In this case, the first document may be comprised mostly of the first few topics; the other document may be mostly about the final few topics (Figure 8).

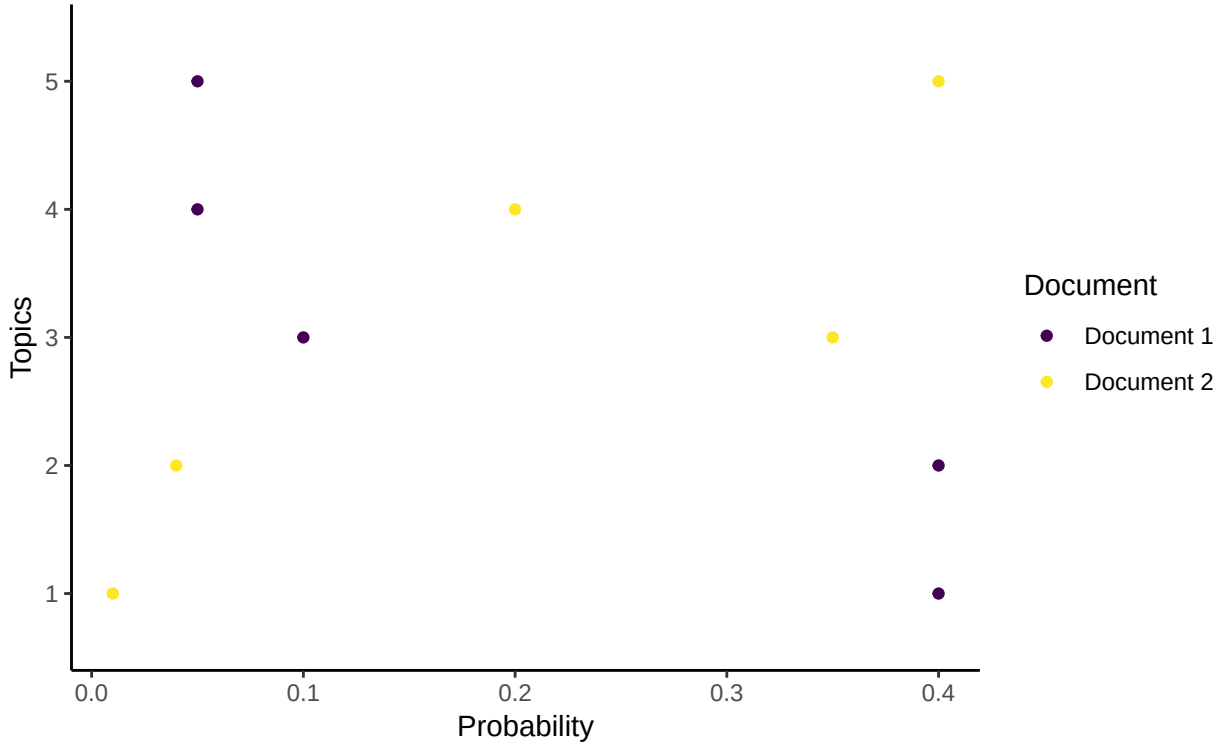


Figure 8: Probability distributions over topics for two documents

For instance, if there were ten terms, then one topic could be defined by giving more weight to terms related to immigration; and some other topic may give more weight to terms related to the economy (Figure 9).

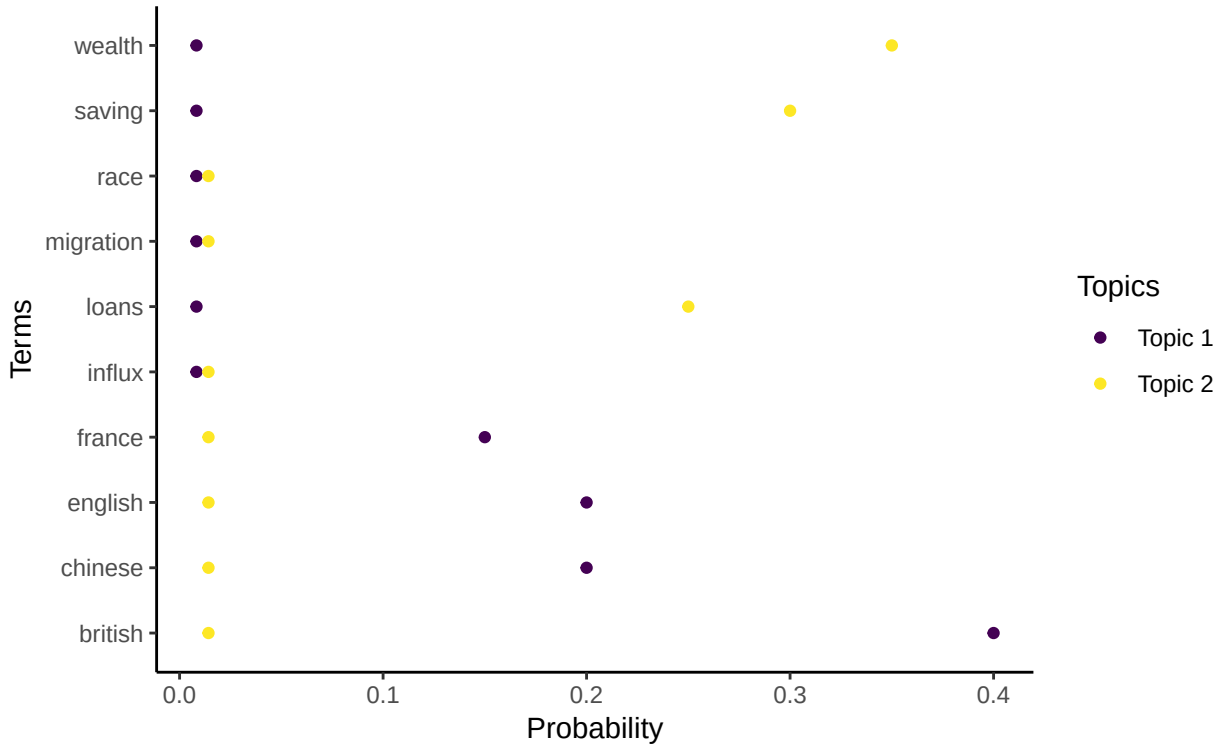


Figure 9: Probability distributions over terms

B.2 Document generation process

Following Blei and Lafferty (2009), Blei (2012) and Griffiths and Steyvers (2004), the process by which a document is generated is more formally considered to be:

1. There are $1, 2, \ldots, k, \ldots, K$ topics and the vocabulary consists of $1, 2, \ldots, V$ terms. For each topic, decide the terms that the topic uses by randomly drawing distributions over the terms. The distribution over the terms for the kth topic is β_k. Typically a topic would be a small number of terms and so the Dirichlet distribution with hyper-parameter $\boldsymbol{\eta}$ is used: $\beta_k \sim \text{Dirichlet}(\boldsymbol{\eta})$, where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_K)$.[4] In practice, a symmetric Dirichlet distribution is usually used, where all elements of $\boldsymbol{\eta}$ are equal.
2. Decide the topics that each document will cover by randomly drawing distributions over the K topics for each of the $1, 2, \ldots, d, \ldots, D$ documents. The topic distributions for the dth document are θ_d, and $\theta_{d,k}$ is the topic distribution for topic k in document d. Again, the Dirichlet distribution with the hyper-parameter $0 < \alpha < 1$ is used here because usually a document would only cover a handful of topics: $\theta_d \sim \text{Dirichlet}(\boldsymbol{\alpha})$. Again, strictly $\boldsymbol{\alpha}$ is vector of length K of hyper-parameters and

[4]The Dirichlet distribution is a variation of the beta distribution that is commonly used as a prior for categorical and multinomial variables. If there are just two categories, then the Dirichlet and the beta distributions are the same. In the special case of a symmetric Dirichlet distribution, where all elements of $\eta = 1$, it is equivalent to a uniform distribution. If $\eta < 1$, then the distribution is sparse and concentrated on a smaller number of the values, and this number decreases as η decreases. A hyper-parameter is a parameter of a prior distribution.

they are usually equal.

3. If there are $1, 2, \ldots, n, \ldots, N$ terms in the dth document, then to choose the nth term, $w_{d,n}$:

    a. Randomly choose a topic for that term n, in that document d, $z_{d,n}$, from the multinomial distribution over topics in that document, $z_{d,n} \sim \text{Multinomial}(\theta_d)$.
    b. Randomly choose a term from the relevant multinomial distribution over the terms for that topic, $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$.

Given this set-up, the joint distribution for the variables is (Blei (2012), p.6):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}, w_{1:D,1:N}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p\left(w_{d,n}|\beta_{1:K}, z_{d,n}\right) \right).$$

Based on this document generation process the analysis problem, discussed next, is to compute a posterior over $\beta_{1:K}$ and $\theta_{1:D}$, given $w_{1:D,1:N}$. This is intractable directly, but can be approximated (Griffiths and Steyvers (2004) and Blei (2012)).

After the documents are created, they are all that we have to analyse. The term usage in each document, $w_{1:D,1:N}$, is observed, but the topics are hidden, or 'latent'. We do not know the topics of each document, nor how terms defined the topics. That is, we do not know the probability distributions of Figures 8 or 9. In a sense we are trying to reverse the document generation process – we have the terms and we would like to discover the topics.

If the earlier process around how the documents were generated is assumed and we observe the terms in each document, then we can obtain estimates of the topics (Steyvers and Griffiths (2006)). The outcomes of the LDA process are probability distributions and these define the topics. Each term will be given a probability of being a member of a particular topic, and each document will be given a probability of being about a particular topic. That is, we are trying to calculate the posterior distribution of the topics given the terms observed in each document (Blei (2012), p. 7):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}|w_{1:D,1:N}) = \frac{p\left(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}, w_{1:D,1:N}\right)}{p(w_{1:D,1:N})}.$$

Gibbs sampling or the variational expectation-maximization algorithm can be used to approximate the posterior. A summary of these approaches is provided next.

B.3 Posterior estimation

Following Steyvers and Griffiths (2006) and Darling (2011), the Gibbs sampling process attempts to find a topic for a particular term in a particular document, given the topics of all other terms for all other documents. Broadly, it does this by first assigning every term in every document to a random topic, specified by Dirichlet priors with $\alpha = \frac{50}{K}$ and $\eta = 0.1$ (Steyvers and Griffiths (2006) recommends $\eta = 0.01$), where α refers to the distribution over topics and η refers to the distribution over terms (Grün and Hornik (2011), p. 7). It then selects a particular term in a particular document and assigns it to a new topic based on the conditional distribution where the topics for all other terms in all documents are taken as given (Grün and Hornik (2011), p. 6):

$$p(z_{d,n} = k | w_{1:D,1:N}, z'_{d,n}) \propto \frac{\lambda'_{n\to k} + \eta}{\lambda'_{\cdot\to k} + V\eta} \frac{\lambda'^{(d)}_{n\to k} + \alpha}{\lambda'^{(d)}_{-i} + K\alpha}$$

where $z'_{d,n}$ refers to all other topic assignments; $\lambda'_{n\to k}$ is a count of how many other times that term has been assigned to topic k; $\lambda'_{\cdot\to k}$ is a count of how many other times that any term has been assigned to topic k; $\lambda'^{(d)}_{n\to k}$ is a count of how many other times that term has been assigned to topic k in that particular document; and $\lambda'^{(d)}_{-i}$ is a count of how many other times that term has been assigned in that document. Once $z_{d,n}$ has been estimated, then estimates for the distribution of words into topics and topics into documents can be backed out.

This conditional distribution assigns topics depending on how often a term has been assigned to that topic previously, and how common the topic is in that document (Steyvers and Griffiths (2006)). The initial random allocation of topics means that the results of early passes through the corpus of document are poor, but given enough time the algorithm converges to an appropriate estimate.

The choice of the number of topics, k, drives the results and must be specified *a priori*. If there is a strong reason for a particular number, then this can be used. Otherwise, one way to choose an appropriate number is to use cross validation. More detail on this process is provided in the next section.

B.4 Selection of number of topics

The choice of the number of topics to use in a topic model has a substantial affect on the results of the model. For instance, in our topic model, choosing a smaller number of topics, such as 10 or 20 results in a model that is not all that useful because the topics are so broad.

There are a variety of diagnostic measures that can guide the selection of the topics, but there is rarely an obvious best choice, especially at a more fine level such as choosing betweeen 60 and 65 topics. We found it useful to try a few quite different measures before settling on 80 topics. This provided a balance between being granular enough to be informative–anything less than 40 topics tended to be too broad– yet still being tractable for our analysis model in a reasonable amount of time. In addition to looking at the topics and how they changed over time, diagnostic measures that we considered include the held-out likelihood, the lower bound, residuals, exclusivity. and semantic coherence (Figures 10).[5]

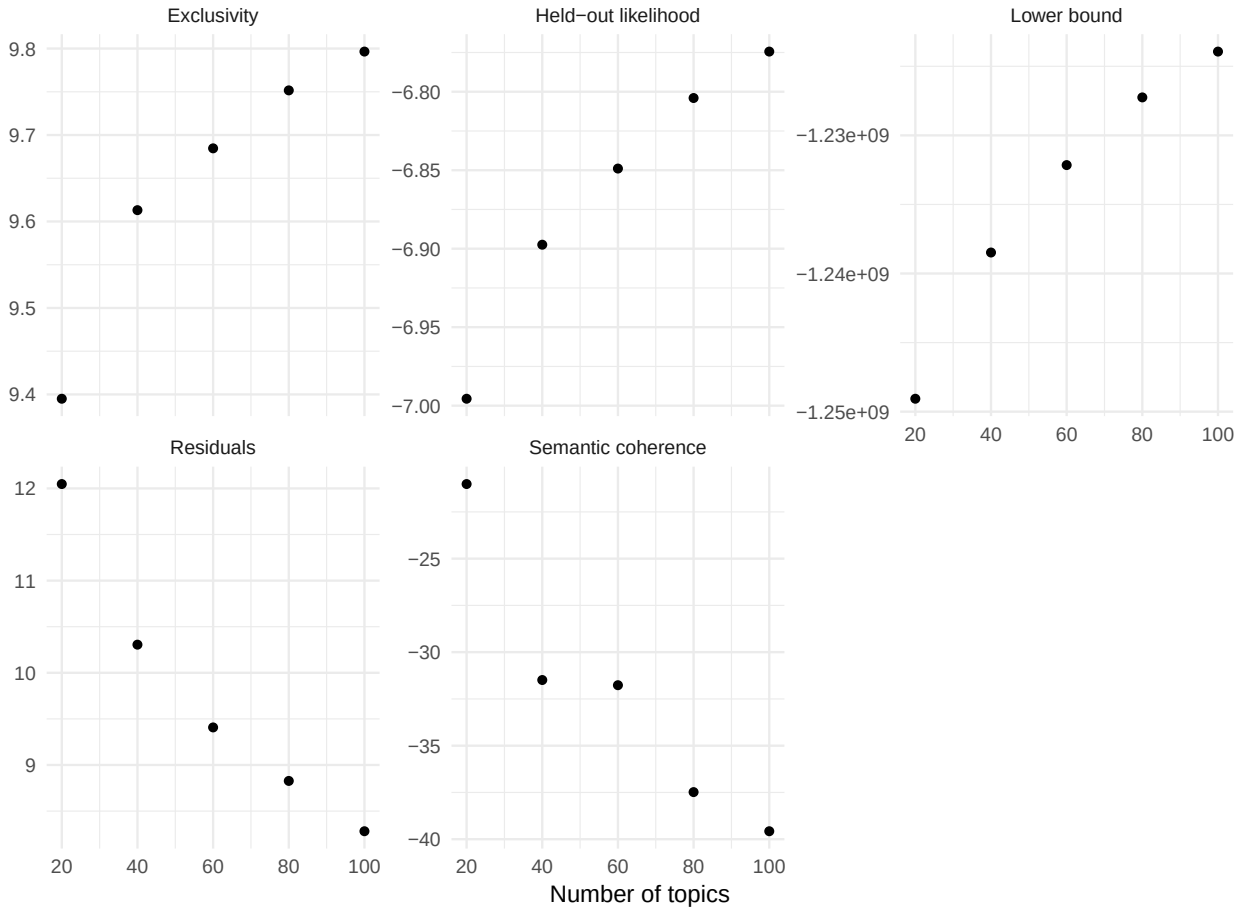


Figure 10: Model diagnostics

Roberts, Stewart and Tingley (2018) provides more detail about the diagnostic tests

[5]The code for creating the figures in this section is based on Silge (2018).

that we use, but we briefly discuss each here. Exclusivity is a measure of how specific words are to particular topics. It looks at the proportion that a word makes up of a particular topic compared with the proportion that word makes up of the other topics. As the number of topics increases we usually expect exclusivity to increase because the topics become more particular. Higher values are better. The held out likelihood as described by Wallach et al. (2009) takes a test/training approach to estimate the probability of held-out documents given the training documents. Higher values are better. The lower bound gives some indication of whether the model may have multiple modes and hence the end result be sensitive to the starting position (Roberts, Stewart and Tingley (2016)). Residuals analysis (Taddy (2012)) compares the theoretical distribution of the variance with the eactual distribution. It is a test for overdispersion of the variance, and if it is found then this can suggest that more topics would be appropriate. Semantic coherence is the trade-off for having topics that are more specific and the subsequent risk that the topics become meaningless. Mimno et al. (2011) define a measure of coherence that is based on ratios of single words compared with pairs of words. The idea is that words that should occur in the one document should be more likely to be in a particular topic than ones that do not occur together. For instance, a topic that has 'wine' and 'cheese' as highly rated words would score better on their measure than another that contained 'cheese' and 'mining'. Lower values are better. Roberts, Stewart and Tingley (2018) recommend examining the tradeoff between exclusitivity and semantic coherence. This suggests that the magnitude of improvement reduces from about 80 topics (Figure 11).

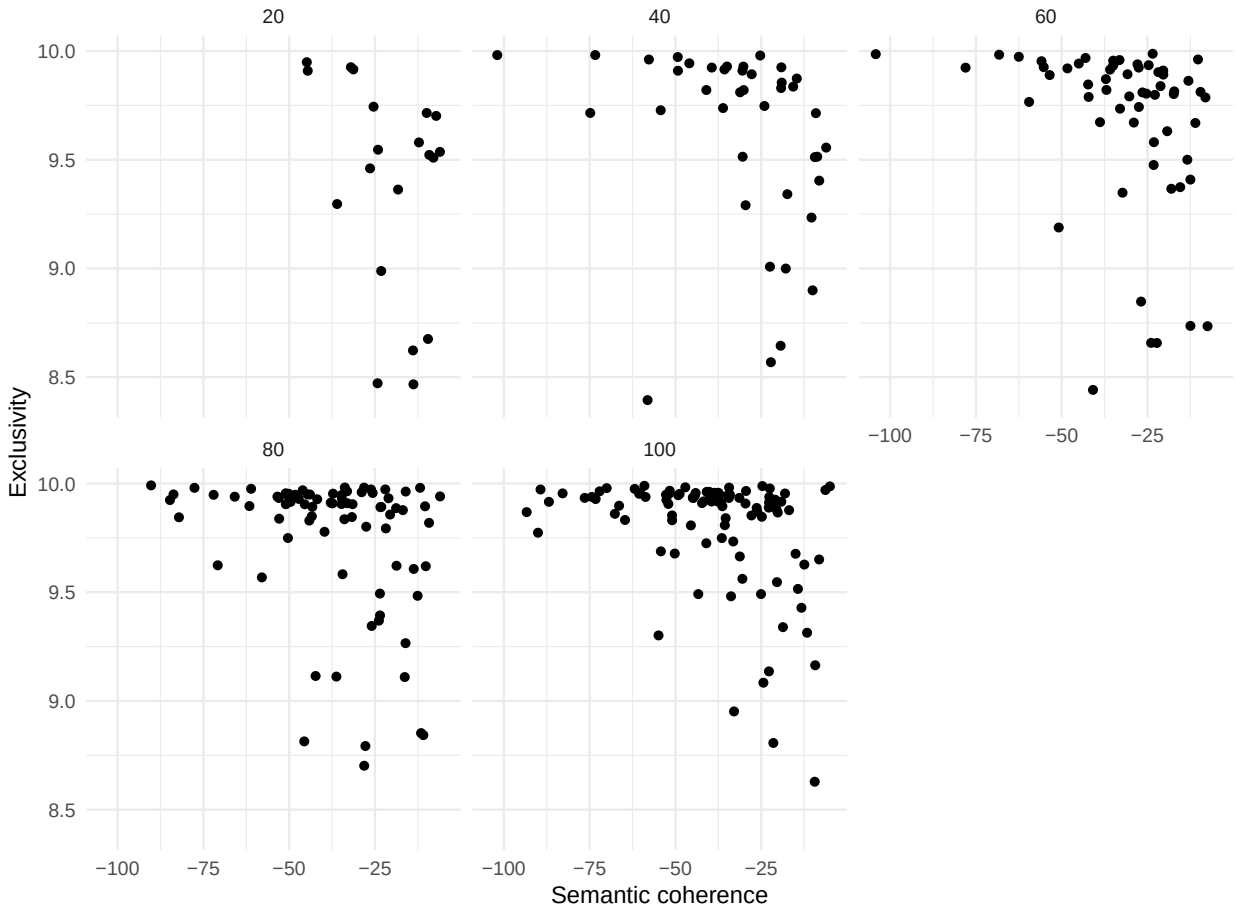


Figure 11: Exclusitivity compared with semantic coherence

C Events

For background on the Premiers' Plan see Copland (1934).
For background on the Gruen tariff cut see Gruen (1975).
Add the graphs and procedures. **[TBD]**

Table 5: Change in governments

government	primeMinister	party	start	end	diedInOffice
Barton	Edmund Barton	Protectionist	1901-01-01	1903-09-24	No
Deakin 1	Alfred Deakin	Protectionist	1903-09-24	1904-04-27	No
Watson	Chris Watson	Labour	1904-04-27	1904-08-18	No
Reid	George Reid	Free Trade	1904-08-18	1905-07-05	No
Deakin 2	Alfred Deakin	Protectionist	1905-07-05	1908-11-13	No
Fisher 1	Andrew Fisher	Labour	1908-11-13	1909-06-02	No
Deakin 3	Alfred Deakin	Commonwealth Liberal	1909-06-02	1910-04-29	No
Fisher 2	Andrew Fisher	Labor	1910-04-29	1913-06-24	No
Cook	Joseph Cook	Commonwealth Liberal	1913-06-24	1914-09-17	No
Fisher 3	Andrew Fisher	Labor	1914-09-17	1915-10-27	No
Hughes	Billy Hughes	Labor, National Labor and Nationalist	1915-10-27	1923-02-09	No
Bruce	Stanley Bruce	Nationalist (Coalition)	1923-02-09	1929-10-22	No
Scullin	James Scullin	Labor	1929-10-22	1932-01-06	No
Lyons	Joseph Lyons	United Australia (Coalition)	1932-01-06	1939-04-07	Yes
Page	Earle Page	Country (Coalition)	1939-04-07	1939-04-26	No
Menzies 1	Robert Menzies	United Australia (Coalition)	1939-04-26	1941-08-28	No
Fadden	Arthur Fadden	Country (Coalition)	1941-08-28	1941-10-07	No
Curtin	John Curtin	Labor	1941-10-07	1945-07-05	Yes
Forde	Frank Forde	Labor	1945-07-06	1945-07-13	No
Chifley	Ben Chifley	Labor	1945-07-13	1949-12-19	No
Menzies 2	Robert Menzies	Liberal (Coalition)	1949-12-19	1966-01-26	No
Holt	Harold Holt	Liberal (Coalition)	1966-01-26	1967-12-19	Yes
McEwen	John McEwen	Country (Coalition)	1967-12-19	1968-01-10	No
Gorton	John Gorton	Liberal (Coalition)	1968-01-10	1971-03-10	No
McMahon	William McMahon	Liberal (Coalition)	1971-03-10	1972-12-05	No
Whitlam	Gough Whitlam	Labor	1972-12-05	1975-11-11	No
Fraser	Malcolm Fraser	Liberal (Coalition)	1975-11-11	1983-03-11	No
Hawke	Bob Hawke	Labor	1983-03-11	1991-12-20	No
Keating	Paul Keating	Labor	1991-12-20	1996-03-11	No
Howard	John Howard	Liberal (Coalition)	1996-03-11	2007-12-03	No
Rudd 1	Kevin Rudd	Labor	2007-12-03	2010-06-24	No
Gillard	Julia Gillard	Labor	2010-06-24	2013-06-27	No
Rudd 2	Kevin Rudd	Labor	2013-06-27	2013-09-18	No
Abbott	Tony Abbott	Liberal (Coalition)	2013-09-18	2015-09-15	No
Turnbull	Malcolm Turnbull	Liberal (Coalition)	2015-09-15	2018-08-24	No
Morrison	Scott Morrison	Liberal (Coalition)	2018-08-24	NA	NA

Table 6: Elections

year	electionDate	electionWinner
1901	1901-03-29	Non-labor
1903	1903-12-16	Non-labor
1906	1906-12-12	Non-labor
1910	1910-04-13	Labor
1913	1913-05-31	Non-labor
1914	1914-09-05	Labor
1917	1917-05-05	Non-labor
1919	1919-12-13	Non-labor
1922	1922-12-16	Non-labor
1925	1925-11-14	Non-labor
1928	1928-11-17	Non-labor
1929	1929-10-12	Labor
1931	1931-12-19	Non-labor
1934	1934-09-15	Non-labor
1937	1937-10-23	Non-labor
1940	1940-09-21	Non-labor
1943	1943-08-21	Labor
1946	1946-09-28	Labor
1949	1949-12-10	Non-labor
1951	1951-08-28	Non-labor
1954	1954-05-29	Non-labor
1955	1955-12-10	Non-labor
1958	1958-11-22	Non-labor
1961	1961-12-09	Non-labor
1963	1963-11-30	Non-labor
1966	1966-11-26	Non-labor
1969	1969-10-25	Non-labor
1972	1972-12-02	Labor
1974	1974-05-18	Labor
1975	1975-12-13	Non-labor
1977	1977-12-10	Non-labor
1980	1980-10-18	Non-labor
1983	1983-03-05	Labor
1984	1984-12-01	Labor
1987	1987-07-11	Labor
1990	1990-03-24	Labor
1993	1993-03-13	Labor
1996	1996-03-02	Non-labor
1998	1998-10-03	Non-labor
2001	2001-11-10	Non-labor
2004	2004-10-09	Non-labor
2007	2007-11-24	Labor
2010	2010-08-21	Labor
2013	2013-09-07	Non-labor
2016	2016-07-02	Non-labor

Table 7: Key events

theDate	event	expected
1901-01-01	Federation	Yes
1902-05-31	Second Boer War ends	Yes
1907-11-08	Harvester case	No
1910-09-10	Australian pound introduced	Yes
1914-07-28	World War I starts	No
1918-11-11	World War I ends	Yes
1929-10-29	Black Tuesday Stock Market Crash	No
1931-06-11	Premiers' Plan agreed	Yes
1932-05-13	Jack Lang dismissed as NSW Premier	Yes
1939-09-01	World War II starts	Yes
1945-09-02	World War II ends	Yes
1949-10-17	Snowy Hydro construction begins	Yes
1956-11-22	Melbourne Olympics	Yes
1962-08-03	Australia enters Vietnam War	Yes
1966-02-14	Decimalisation	Yes
1972-12-02	Australia exits Vietnam War	Yes
1973-01-18	Gruen tariff cut	Yes
1973-10-20	White Australian Policy ended	Yes
1975-11-11	The Dismissal	No
1983-12-12	Australian dollar is floated	Yes
1984-02-01	Medicare establised	Yes
1987-10-19	Black Monday Stock Market Crash	No
1990-08-27	State Bank of Victoria collapse	No
1991-02-10	State Bank of South Australia collapse	No
1991-02-10	First Gulf War begins (FIX DATE)	No
1992-06-03	Mabo decision	No
1996-03-28	Port Arthur massacre	No
1996-12-23	Wik decision	No
1999-09-20	INTERFET deployment begins	No
2000-07-01	GST introduced	Yes
2000-09-15	Sydney Olympics	Yes
2001-09-11	9/11 attack	No
2002-10-12	Bali bombings	No
2008-09-15	Lehman Brothers bankruptcy	No

References

Baumgartner, Frank R. and Bryan D. Jones. 1993. *Agendas and Instability in American Politics*. University of Chicago Press.

Beelen, Kaspar, Timothy Alberdingk Thim, Christopher Cochrane, Kees Halvemaan, Graeme Hirst, Michael Kimmins, Sander Lijbrink, Maarten Marx, Nona Naderi, Ludovic Rheault, Roman Polyanovsky and Tanya Whyte. 2017. "Digitization of the Canadian Parliamentary Debates." *Canadian Journal of Political Science* pp. 1–16.

Benoit, Kenneth. 2018. *quanteda: Quantitative Analysis of Textual Data*. R package version 1.3.4.
URL: *http://quanteda.io*

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55(4):77–84.

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

Blei, David M and John D Lafferty. 2007. "A Correlated Topic Model of Science." *The Annals of Applied Statistics* 1(1):17–35.

Blei, David M and John D Lafferty. 2009. Topic Models. In *Text Mining*. Chapman and Hall/CRC pp. 101–124.

Boulus, Paul. 2013. *Ask Keith for title*. ANU Honours thesis.

Copland, Douglas. 1934. "The Premiers' Plan in Australia: An Experiment in Economic Adjustment." *International Affairs (Royal Institute of International Affairs 1931-1939)* 13(1):79–92.

Curran, B., K. Higham, E. Ortiz and D. Vasques Filho. 2017. "Look Who's Talking: Bipartite Networks as Representations of a Topic Model of New Zealand Parliamentary Speeches." *ArXiv e-prints* .

Darling, William M. 2011. A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. pp. 642–647.

Dimitruk, Kara. 2018. "I Intend Therefore to Prorogue: The Effects of Political Conflict and the Glorious Revolution in Parliament, 1660–1702." *European Review of Economic History* 22(3):261–297.

Dowding, Keith, Andrew Hindmoor, Richard Iles and Peter John. 2010. "Policy Agendas in Australian Politics: The Governor-General's Speeches, 1945–2008." *Australian Journal of Political Science* 45(4):533–557.

Duthie, Rory, Katarzyna Budzynska and Chris Reed. 2016. Mining Ethos in Political Debate. In *Computational Models of Argument*, ed. P Baroni, TF Gordon, T Scheffler and M Stede. Vol. 287 pp. 299–310.

Edwards, Cecilia. 2016. "The Political Consequences of Hansard Editorial Policies: The Case for Greater Transparency." *Australasian Parliamentary Review* 31(2):145–160.

Feinerer, Ingo and Kurt Hornik. 2018. *tm: Text Mining Package.* R package version 0.7-5.
URL: *https://CRAN.R-project.org/package=tm*

Fraussen, B, T Graham and D Halpin. forthcoming. "Assessing The Prominence Of Interest Groups In Parliament: A Supervised Machine Learning Approach." *Journal of Legislative Studies* .

Gagolewski, Marek. 2018. *R Package stringi: Character String Processing Facilities*.
URL: *http://www.gagolewski.com/software/stringi/*

Gans, Joshua and Andrew Leigh. 2012. "How Partisan Is The Press? Multiple Measures Of Media Slant." *The Economic Record* 88(280):127–147.

Gentzkow, Matthew, Jesse M. Shapiro and Matt Taddy. 2018. Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. Technical report Voldemort's University.
URL: *http://web.stanford.edu/ gentzkow/research/politext.pdf*

Graham, Ruth. 2016. Withdraw and Apologise: A Diachronic Study of Unparliamentary Language in the New Zealand Parliament, 1890–1950 PhD thesis.

Griffiths, Thomas and Mark Steyvers. 2004. "Finding Scientific Topics." *PNAS* 101:5228–5235.

Grolemund, Garrett and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40(3):1–25.
URL: *http://www.jstatsoft.org/v40/i03/*

Gruen, F. H. 1975. "The 25% Tariff Cut; Was It a Mistake?" *The Australian Quarterly* 47(2):7–20.

Grün, Bettina and Kurt Hornik. 2011. "topicmodels: An R Package for Fitting Topic Models." *Journal of Statistical Software* 40(13):1–30.

Izrailev, Sergei. 2014. *tictoc: Functions for Timing R Scripts*. R package version 1.0.
URL: *https://CRAN.R-project.org/package=tictoc*

Mimno, David, Edmund Talley, Miriam Leenders, Hanna M Wallach and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK pp. 262–272.

Mollin, Sandra. 2008. "The Hansard hazard: Gauging the Accuracy of British Parliamentary Transcripts." *Corpora* 2(2):187–210.

Ooms, Jeroen. 2017. *hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker*. R package version 2.9.
URL: *https://CRAN.R-project.org/package=hunspell*

Ooms, Jeroen. 2018. *pdftools: Text Extraction, Rendering and Converting of PDF Documents*. R package version 1.8.
URL: *https://CRAN.R-project.org/package=pdftools*

Peterson, Andrew and Arthur Spirling. 2018. "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems." *Political Analysis* 26(1):120–128.

Plummer, Martyn. 2018. *rjags: Bayesian Graphical Models using MCMC*. R package version 4-7.
URL: *https://CRAN.R-project.org/package=rjags*

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *https://www.R-project.org/*

Rasiah, Parameswary. 2010. "A Framework for the Systematic Analysis of Evasion in Parliamentary Discourse." *Journal of Pragmatics* 42:664–680.

Rheault, Ludovic and Christopher Cochrane. 2018. Word Embeddings for the Estimation of Ideological Placement in Parliamentary Corpora. In *PolMeth 2018*. Provo, UT: Society for Political Methodology.

Roberts, Margaret E, Brandon M Stewart and Dustin Tingley. 2016. "Navigating the Local Modes of Big Data." *Computational Social Science* 51.

Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. 2018. *stm: R Package for Structural Topic Models*. R package version 1.3.3.
URL: *http://www.structuraltopicmodel.com*

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* 111(515):988–1003.

Salisbury, Christopher. 2011. "'Mr Speaker, I Withdraw...': Standards Of (mis)behaviour In The Queensland, Western Australian And Commonwealth Parliaments Compared Via Online Hansard." *Australasian Parliamentary Review* 26(1):166–177.

Silge, Julia. 2018. *Training, Evaluating, And Interpreting Topic Models*. Last accessed: 2018-11-06.
URL: *https://juliasilge.com/blog/evaluating-stm/*

Silge, Julia and David Robinson. 2016. "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS* 1(3).
URL: *http://dx.doi.org/10.21105/joss.00037*

Steyvers, Mark and Tom Griffiths. 2006. Probabilistic Topic Models. In *Latent Semantic Analysis: A Road to Meaning*, ed. T. Landauer, D McNamara, S. Dennis and W. Kintsch.

Taddy, Matt. 2015. "Distributed multinomial regression." *The Annals of Applied Statistics* 9(3):1394–1414.

Taddy, Matthew. 2012. On Estimation and Selection for Topic Models. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*. La Palma, Canary Islands pp. 1184–1193.

Van Zanden, Jan Luiten, Eltjo Buringh and Maarten Bosker. 2012. "The Rise and Decline of European Parliaments, 1188–1789." *The Economic History Review* 65(3):835–861.

Vaughan, Davis and Matt Dancho. 2018. *furrr: Apply Mapping Functions in Parallel using Futures*. R package version 0.1.0.9002.
URL: *https://github.com/DavisVaughan/furrr*

Wallach, Hanna M, Iain Murray, Ruslan Salakhutdinov and David Mimno. 2009. Evaluation Methods for Topic Models. In *Proceedings of the 26th International Conference on Machine Learning*. Montreal, Canada.

Wickham, Hadley. 2017. *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.
URL: *https://CRAN.R-project.org/package=tidyverse*