The Effect of Events on Discussion in the Australian Federal Parliament (1901–2017)

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

25 October 2018

Presentation at the Parliamentary Library, Canberra, Australia.

Summary

Approach

- Create a dataset of what was said in the Australian Federal Parliament from 1901 through to 2017 based on available public records.
- Use a correlated topic model to reduce dimemsionality.
- Analyse the effect of various events using a Bayesian hierarchical Dirichlet model.

Findings

- 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.
- The effect of events is much more pronounced in the second half of our sample, and especially in the past two decades.

Data

Hansard

- 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. It's not vertatim, but it's pretty close.
- Hansard records have been used in other counties but not really at scale in Australia (although others, such as Patrick Leslie, also use them).
- Daily PDFs are available Hansard records available online as PDFs and these are considered the official release.
- There are 14,551 days of publicly available Hansard records across both chambers of the Australian Federal Parliament.

Example PDF page

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?

PDF parsing

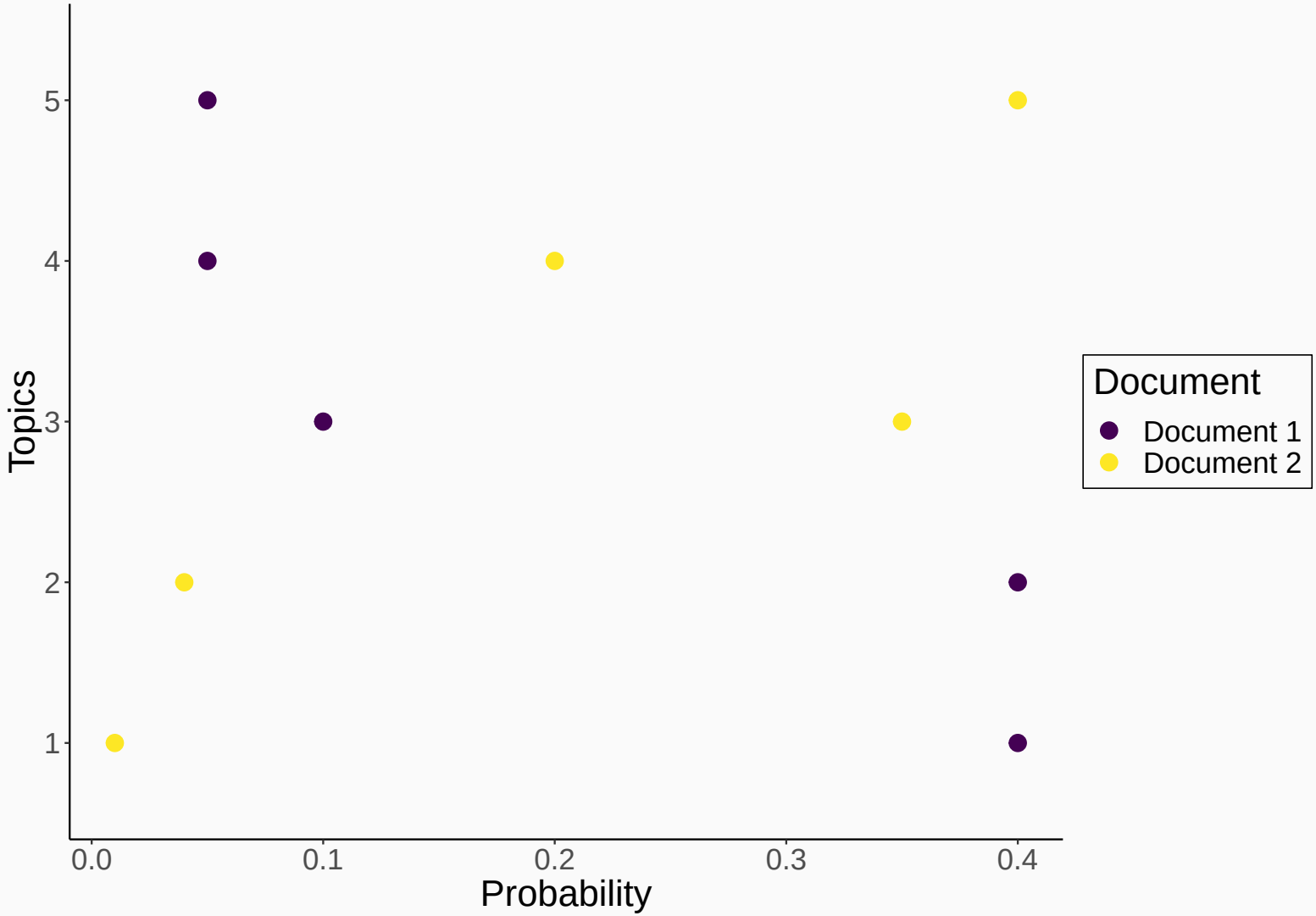
- We use scripts written in R to convert the PDFs into daily text records.
- Some error is introduced at this stage:
  - because many of the records are in a two-column format that need to be separated; and
  - the PDF parsing is not always accurate especially for older records e.g. 'the' is often parsed as 'thc'.
- We 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.

Topic Model

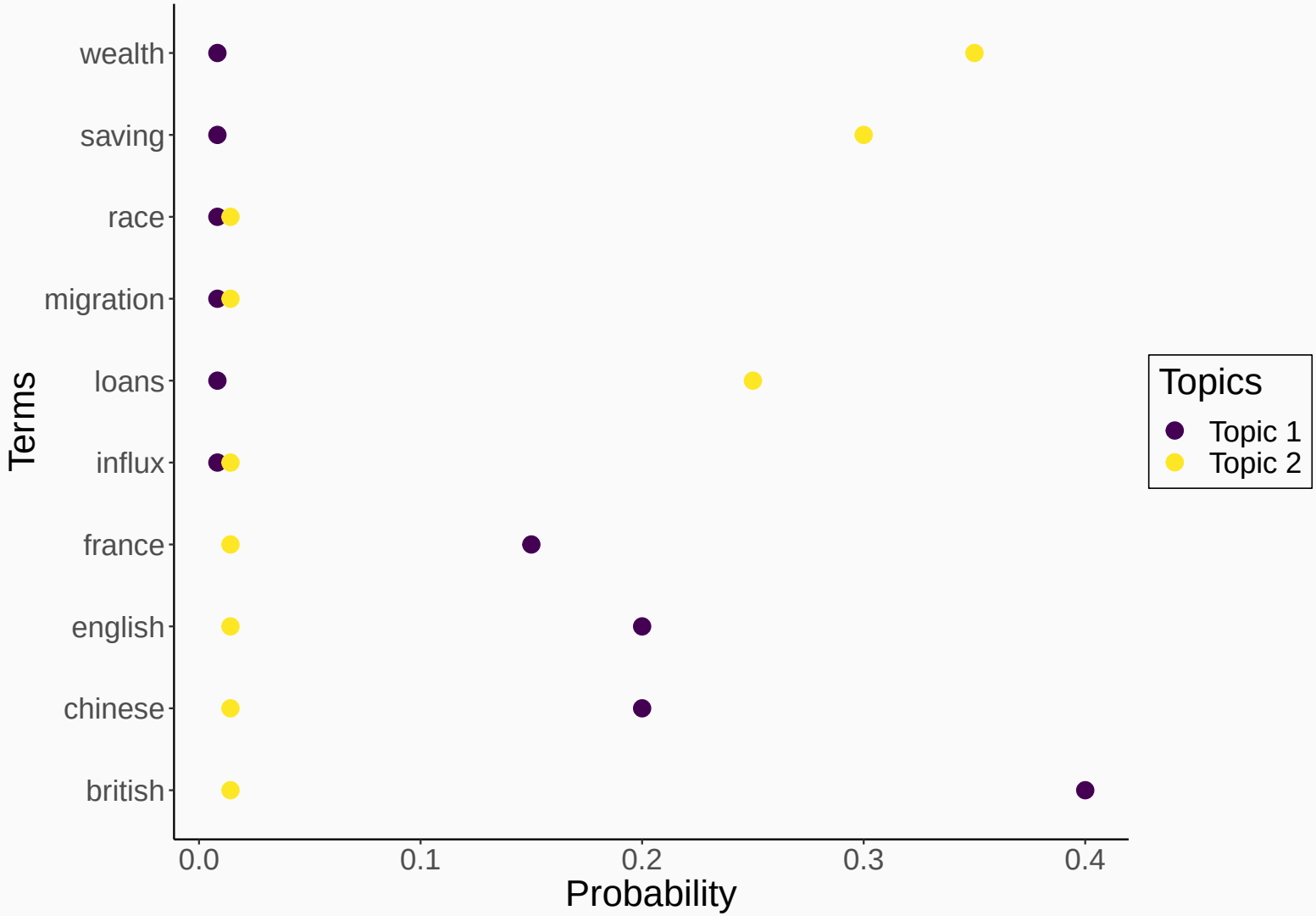
Latent Dirichlet Allocation - Overview

- Each document is assumed to be from a speaker who decides 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 – this is an unsupervised machine learning method.

Latent Dirichlet Allocation - Topic distribution for documents



Latent Dirichlet Allocation - Term distribution over topics



Latent Dirichlet Allocation - Data generation process

1. There are K topics and the vocabulary consists of V terms. For each topic, decide the terms by randomly drawing distributions over the terms. The distribution over the terms for the kth topic is β_k. Use the Dirichlet distribution with hyper-parameter $\boldsymbol{\eta}$: $\beta_k \sim \text{Dirichlet}(\boldsymbol{\eta})$, where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_K)$.

Latent Dirichlet Allocation - Data generation process (cont.)

2. Decide the topics that each document will cover by randomly drawing distributions over the K topics for each of the 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: $\theta_d \sim \text{Dirichlet}(\boldsymbol{\alpha})$.

Latent Dirichlet Allocation - Data generation process (cont.)

3. If there are N terms in the dth document, then to choose the nth term, $w_{d,n}$:
   - 3.1 Randomly choose a topic for that term n, in that document d, $z_{d,n}$, from the multinomial distribution over topics in that document, $z_{d,n} \sim \text{Multinomial}(\theta_d)$.
   - 3.2 Randomly choose a term from the relevant multinomial distribution over the terms for that topic, $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$.

Latent Dirichlet Allocation - Data generation process (cont.)

Given this set-up, we have a joint distribution for the variables and 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.

Latent Dirichlet Allocation - Analysis problem

- 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. In a sense we are trying to reverse the document generation process.
- 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.

Latent Dirichlet Allocation - Analysis problem (cont.)

- 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.
- The choice of the number of topics, k, drives the results and must be specified *a priori*.

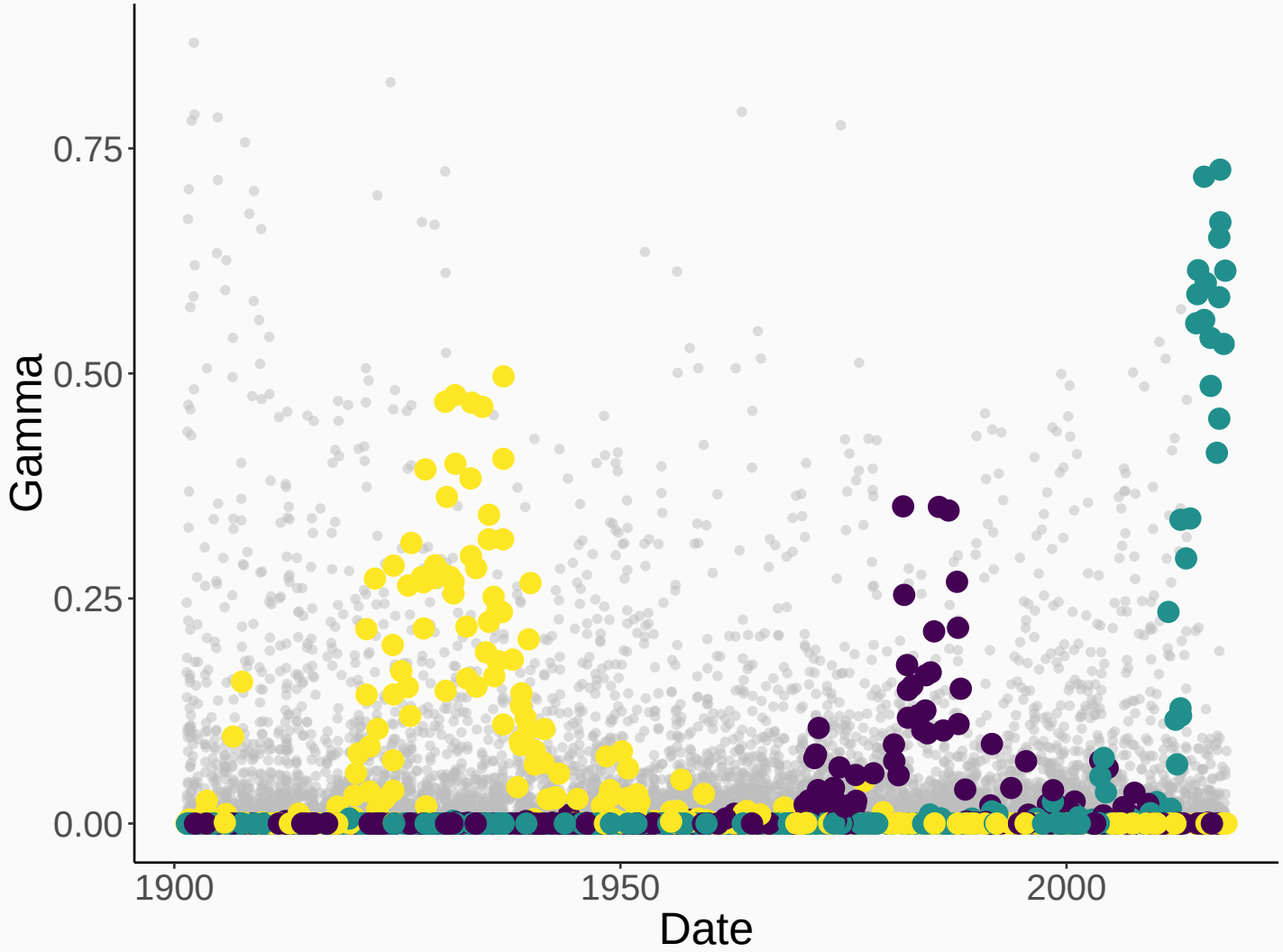
Correlated Topic Model

Modification of LDA - 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

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

Essentially it swaps the Dirichlet for the Logistic Multivariate Normal. This sounds easy, but it's hard to implement.

Correlated Topic Model - Example output



Structural Topic Model

The Structural Topic Model approach then adds a covariate to μ which allows consideration of additional information.

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

Again, sounds easy, but hard to implement, and they have a very nice solutions algorithm.

Why not use STM?

STM is great, but it is specific to various contexts and ours is slightly outside them:

- No way to specify more complicated auto-correlated functional forms of the effects of events over time.
- There is no way to implement partial pooling across groups of similar documents.
- 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. In an ideal world we could modify the STM code, but instead we build a similar, but different, analysis model.

Analysis model

Analysis model - Overview

- We use the estimated topic distributions from the CTM described in the previous section as an input into a Bayesian hierarchical Dirichlet regression framework.
- This 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.

Analysis model - Set-up

- 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.

Analysis model - Daily distribution

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 $\mu_{s[d],1:P}$:

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

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.

Analysis model - Mean parameters

The mean parameters $\mu_{s,p}$ are modelled on the log scale as:

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

- $\alpha_{g[s],p}$ 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.

Analysis model - Government effects

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

Analysis model - Election effects

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:

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

We use non-informative priors.

Analysis model - Time effects

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.

Analysis model - Sitting period effects

The sitting period-specific random effect $\delta_{s,p}$ allows the topic distributions in some sitting periods to be different than expected based on government and election effects:

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

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.

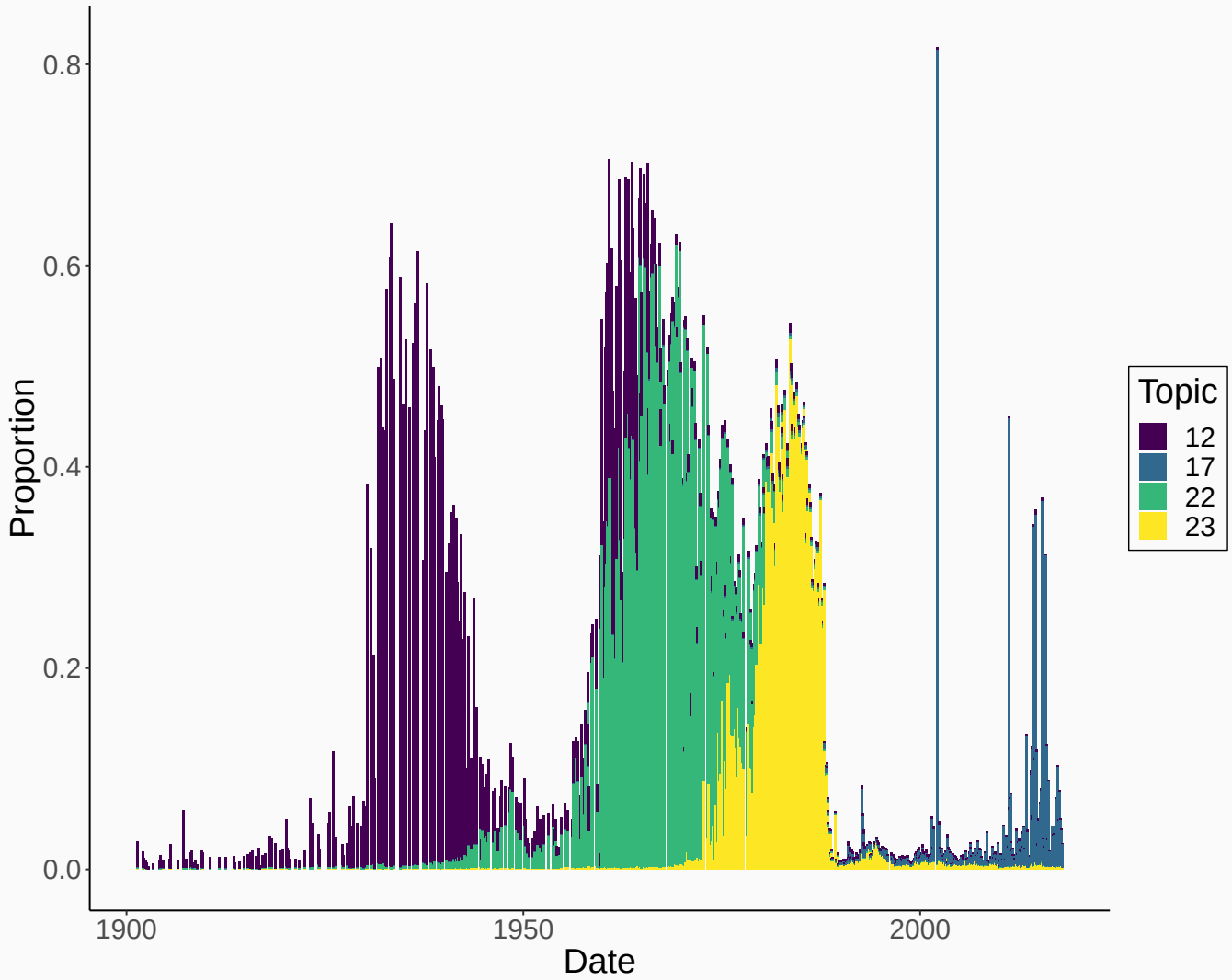
Results

Topic modelling

We choose 40 topics because it is the minimum needed to have meaningful results.

topic	terms
2	constitution, parliament, rights, powers, constitutional
4	bank, money, country, per_cent, budget
6	tax, income, taxation, treasurer, per_cent
8	department, service, officers, office, estimates
10	court, law, royal, evidence, attorneygeneral
12	war, defence, country, ill, soldiers

Analysis model - Estimates by sitting period



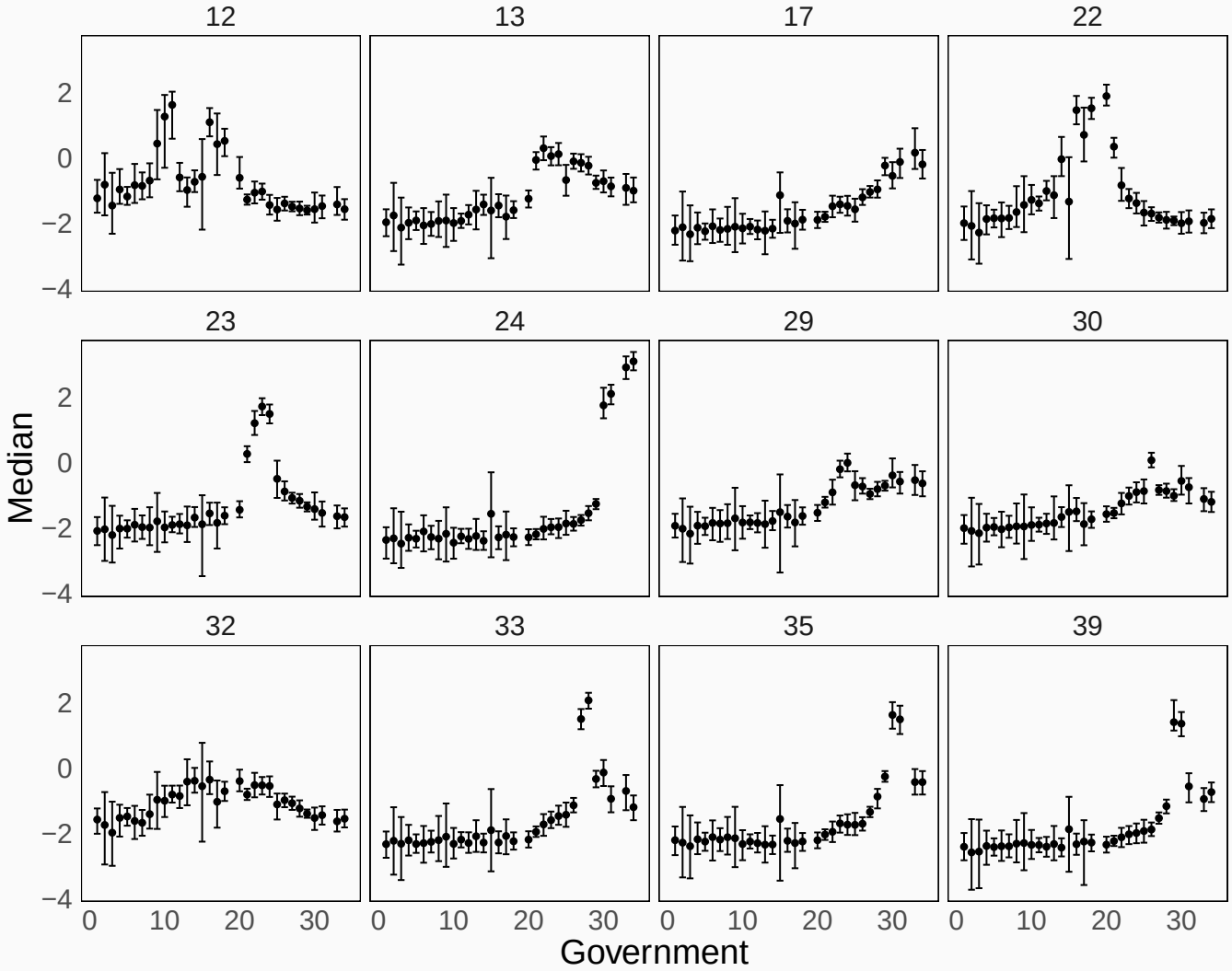
Analysis model - Significantly different elections

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

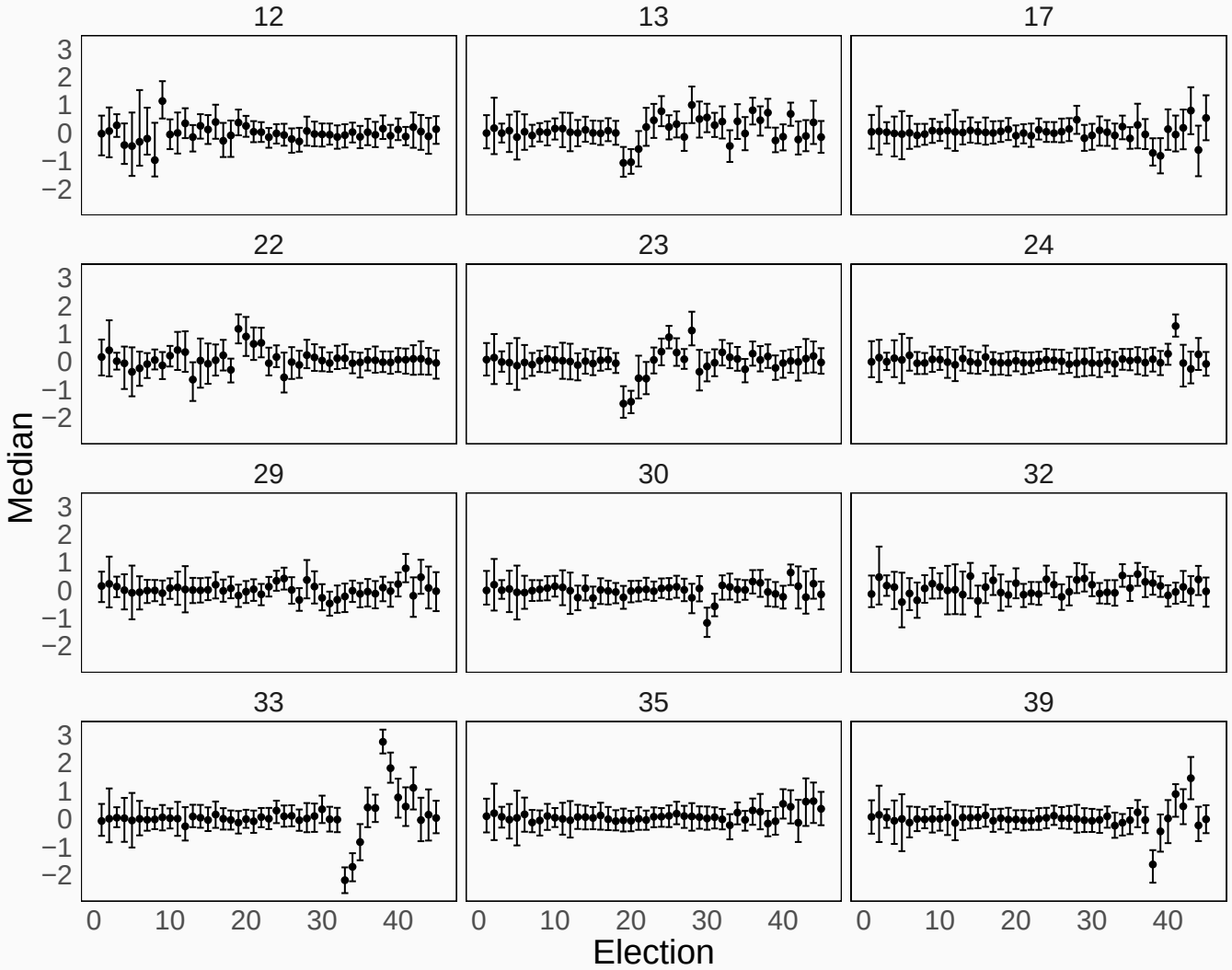
Analysis model - Significantly different governments

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

Analysis model - All governments



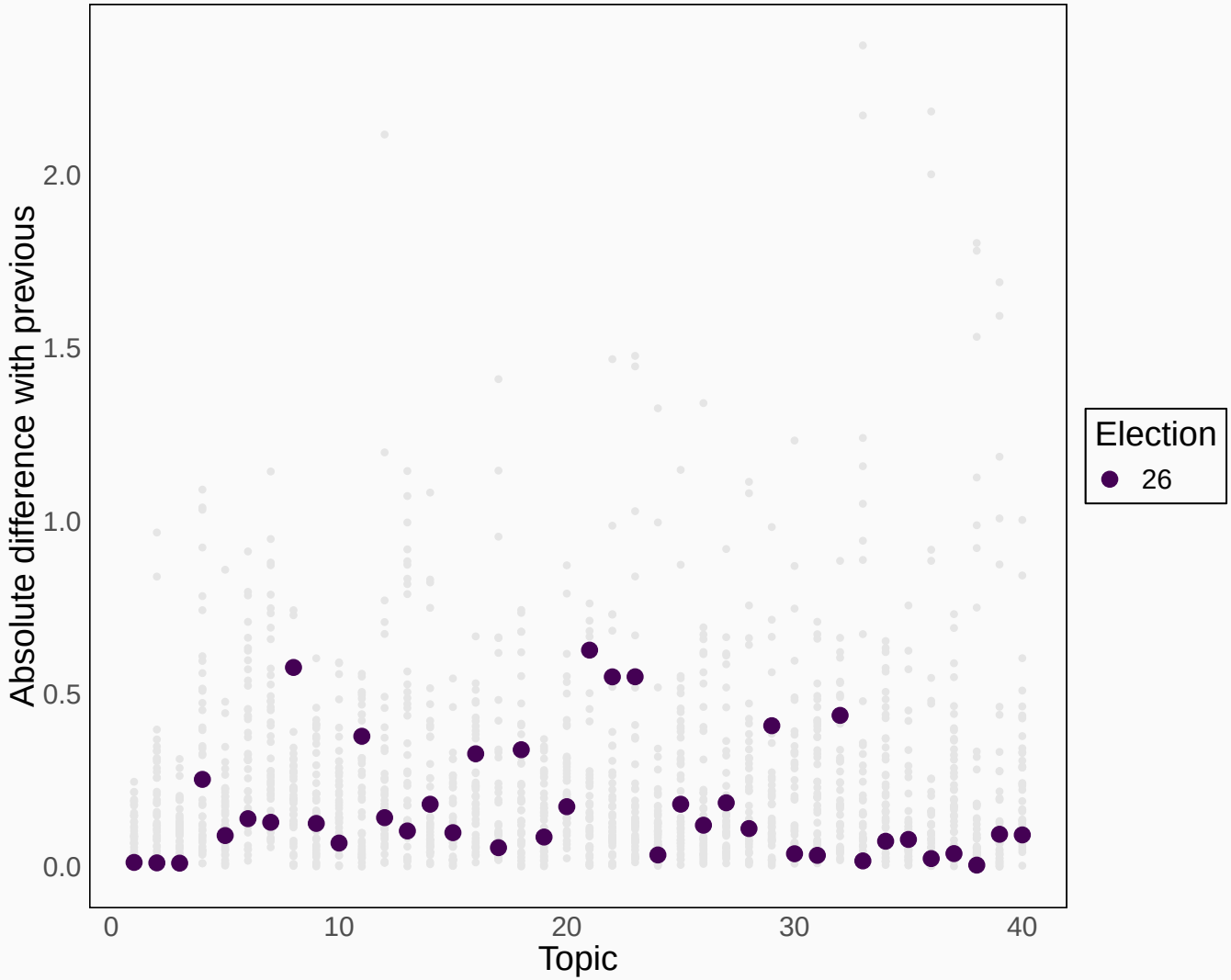
Analysis model - All elections



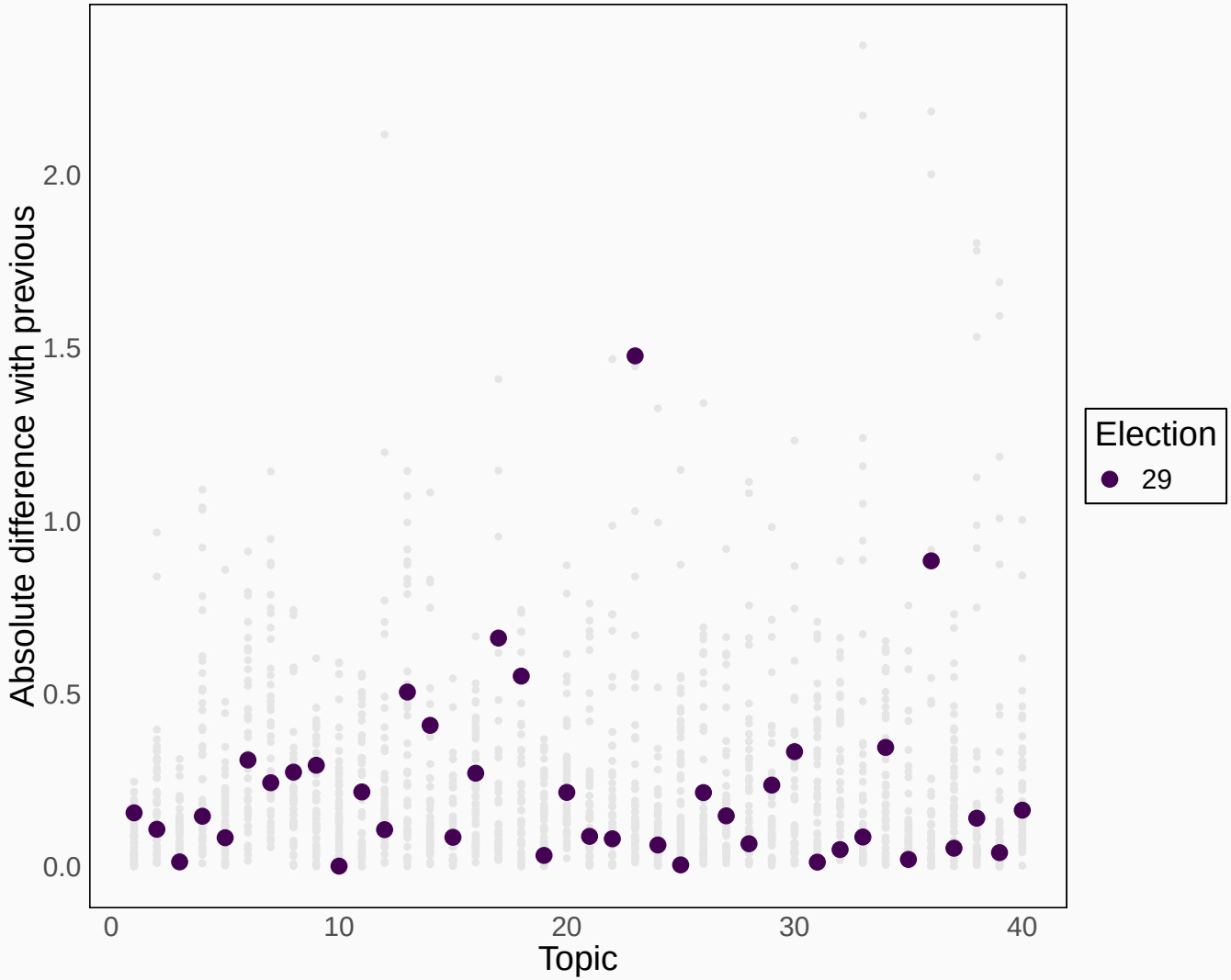
Analysis model - Events

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

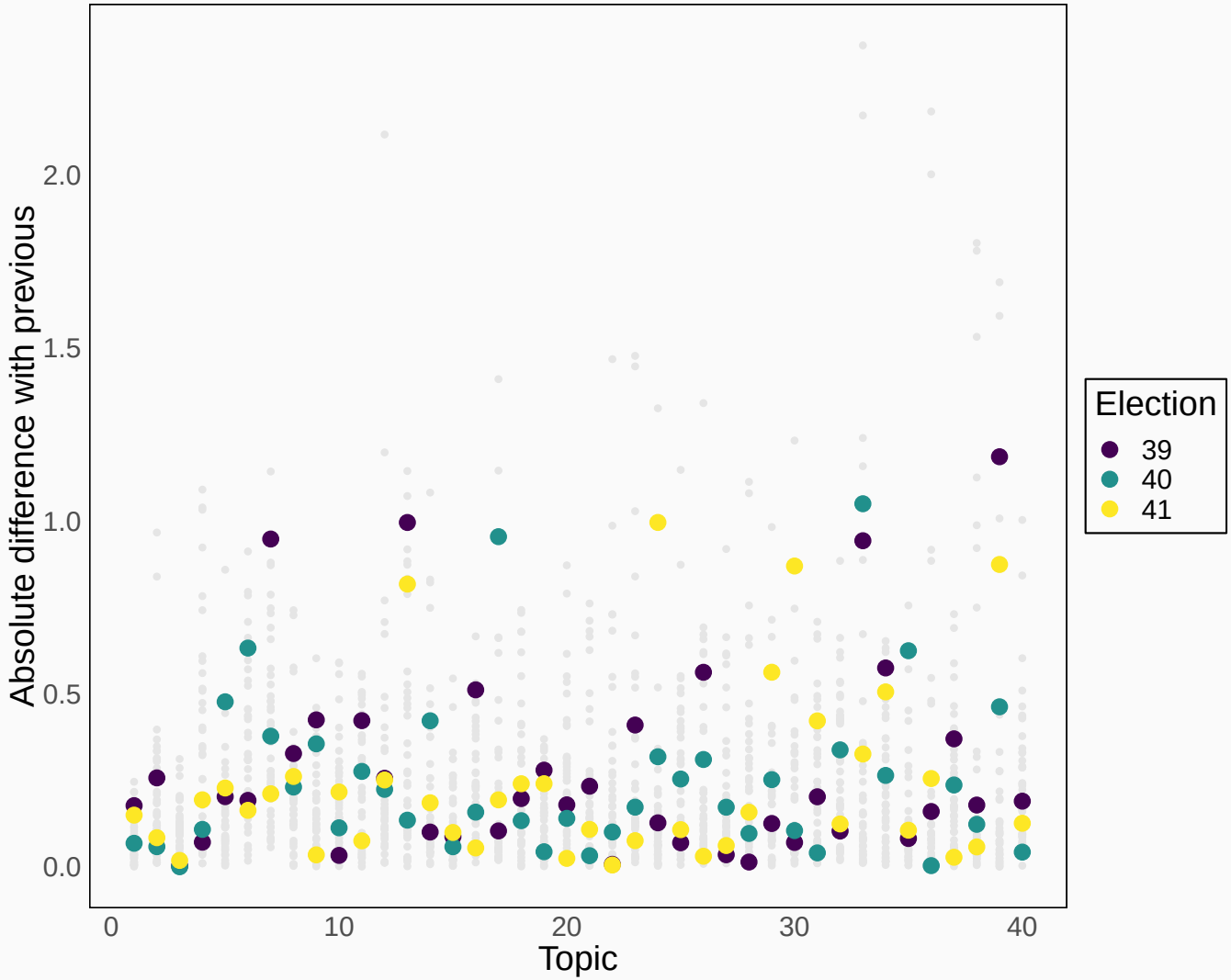
Menzies-Holt



Whitlam re-election



Howard instability



Issues/tradeoffs

Issues/tradeoffs

- Even after cleaning the dataset remains imperfect and is more fit-for-purpose than of broad applicability.
- Topic interpretation is difficult.
- Inputs have to be fine-tuned.
- Assuming government effects are is constant across the whole period.
- Assuming the effect of elections is monotonically decreasing across days since election.
- Identification of unusual periods could also be improved.
- Propogation of uncertainty is inappropriate.

Current/future work

Current/future work

Improve data

- Trying to get website data instead of needing to parse PDFs.

Current work

- In-group and out-group identity: which is more important?
- Senate/HoR: just how different are they?
- Methods assessment: how is your cleaning impacting your results?
- States/Commonwealth: where does policy come from?

Questions?

- rohan.alexander@anu.edu.au
- @RohanAlexander
- rohanalexander.com

COMMONWEALTH OF AUSTRALIA

PARLIAMENTARY DEBATES

House of Representatives

Official Hansard

No. 3, 2017

Monday, 27 February 2017

FORTY-FIFTH PARLIAMENT