#KillTheBill: Evaluating UK parliamentary debates on protest with structural topic modelling

Protests against the Industrial Relations Bill in Trafalgar Square, January 1971

Introduction

In 1971, workers went on strike and marched through London chanting 'kill the bill.' They were protesting the introduction of the Industrial Relations Bill, which would restrict the right to protest. In 1994, 50,000 people marched under the banner 'kill the bill', in protest against the Criminal Justice and Public Order Bill; a bill which gave police more powers to stop and search, and to clamp down on 'anti-social behaviour'. In May 2021, #killthebill trends on Twitter, as people take to the streets once again, against the Police, Crime and Sentencing Bill; a bill which will criminalise protest that may be deemed a 'serious annoyance' (UK Government, 2021).

Protest and civil unrest are widely understood to occur in local and global cycles; triggered by significant events such as economic recessions, political change, or cuts to public services (Tarrow, 1998; Jenkins et al., 2008; Sombatpoonsiri, 2015; Maroto et al., 2019). In the UK, the most recent 'cycles of contention' have peaked in the 1980s, in response to widespread unemployment under Thatcher's government, and in the 2010s, peaking after the global recession in 2008 (Clement, 2019). The COVID-19 crisis has seen a new surge of protest and social movements, in spite of the risks and restrictions brought about by the pandemic (Gerbaudo, 2020). Activists and protest organisers believe that the protests seen through the pandemic mark the beginning of a new period of dissent, unrest and protest

movements, in response to the economic hardship and social injustices exacerbated by the crisis (Chessum, 2021).

Cycles of protest are often followed by a form of clampdown, often by changes in legislation to either restrict the legality of protest, or to amplify the powers of police and the state to suppress protesters (Fernandez, 2009). In the UK, this has been notable in changes to the Public Order Act by Thatcher in 1986, which followed the Miners' Strikes, and most recently with the Police Crime and Sentencing Bill, which activists see as a direct response to protests by Extinction Rebellion and Black Lives Matter (Extinction Rebellion, 2021; Mead, 2021).

The right to protest is a fundamental human right (UN General Assembly, 1948), but in practice, how this right is interpreted and protected in legislation can vary depending on the party in power (della Porta, 1998). Protests as events are critical in the building of movements (della Porta, 2008), and scholars have noted the potential individual and group empowerment of attending a protest march, and expressing demands along with many thousands of others (Opp, 2009; Gerbaudo, 2020). Protests, strikes and political actions are sites of solidarity, power, discussion and democracy, and can therefore be feared by those who govern (Garland, 2002).

This paper explores the latent themes that have run through politicians' speeches about protest in the UK House of Commons over the last 50 years. The transcripts from any debates discussing protest, and from changes to legislation covering protest, are analysed using topic modelling via the **stm** package in R. Through uncovering the latent patterns and themes in parliamentary discourse, the paper seeks to understand the ways in which politicians have framed protest and police power in the UK. In doing so, it seeks to answer the following substantive questions:

1. What latent themes lie underneath politicians' discussion of political protest in House of Commons debates?
2. How does party affiliation affect politicians' framing of political protest?

Background

Power and Social Order

Democratic states are invested in framing policing in terms of maintaining social harmony. By portraying the primary aims of policing as combatting crime and violence to protect social order, governments can hide the true nature of policing: acting as an arm of the state to exercise power and control, and protecting state property and institutions (Waddington, 1987). As industrial capitalism grew in the UK, the role of the police shifted from a very broad remit, to ensure social security, towards the protection of property and institutions. Policing is used to fabricate and recreate a hegemonic social order; an order organised around private property ownership and exploitation (Neocleous, 2000). According to Butler, capitalist systems have an innate need for authoritarianism, and as capitalism has grown, tougher police control has followed (Butler, 2006). Most recently, Maroto sees contemporary policing in the UK as a means for the state to control expression of alternative systems, while externally framing its aims as combatting crime (Maroto et al., 2019).

The UK Conservative Party have been critiqued for using the alleged need to preserve social order to impose restrictions on the right to protest. Ford and Novitz review their consultation ahead of the Trade Union Bill in 2015, and show that the government used the alleged threat to social order to propose much more restrictive regulations on trade unions' right to organise protest action (Ford and Novitz, 2015). The Labour Party have a very different history: as a party born from the working labour movement, they have historically had strong relationships with trade unions and therefore supported organised protest (Ludlam, 2000; Kogan, 2019). Since the Political Fund Ballots of the mid 1980s (Stewart, 2011), and the birth of New Labour in the 1990s (McIlroy, 1998), the relationship with the unions has been more of a what Minkin characterised as a 'contentious alliance' (Minkin, 1991), and support for protest has been more cautious (Kogan, 2019).

Relating hypotheses

H1 Restrictions on protest will be framed around the need to preserve social order.

H2 Conservative Party speeches will be more associated with latent themes around the protection of social order from the threat of crime.

H3 Labour Party speeches will be more likely to frame protest in terms of democratic rights and freedoms.

Framing Protest as Crime

Going back as far as Plato, there has been an impetus to fear the mob; people lost in the crowd are thought to have lost their rationality, swept up in the power of the group (McClelland, 1989). In his study of U.S. police training materials and magazines, Schweingruber found examples of 'mob psychology': a theory of crowd behaviour that justifies the use of force in managing protest. Once a crowd has become a mob, the individuals get swept up in a frenzy, encouraged by the experience of being part of a larger group. This senselessness then renders the mob a danger to society, making it the duty of the police to respond with necessary force (Schweingruber, 2000).

The state and the media use the image of the police in order to construct protest as criminal mob, as can be seen in this headline from the Sun: 'Police were *forced* to let off smoke bombs at a demo as *30,000 protesters descended on* Manchester' (The Sun, 2017, emphasis my own). In a Durkheimian view of punishment, police crackdown is a symbolic construction for the onlookers, to demonstrate control (Durkheim in Garland, 1991). When we see the police intervening in protest, it is an indication that the activities happening must be criminal. As Maroto shows in several global case studies, this has served to present protests as a matter of legality, rather than as demonstrations of a political viewpoint or movement (Maroto, 2019). Framing the mob as depoliticised, criminal, and dangerous serves to create an 'us and them' with the general public, encouraging perception of protest as a criminal threat to public order (Maroto, 2019). In the UK this is explicitly manufactured, as has been demonstrated by Mawby's analysis of the Metropolitan Police's 'impression management' (Mawby, 2014), and Lee and McGovern's qualitative study with police officers, looking at their visual strategies for manufacturing public confidence (Lee and McGovern, 2013).

Relating hypotheses

H4 Protest will be framed in terms of criminality, rather than in terms of politics.

Demonisation and the Law-Abiding Majority

The demonisation of protesters as a criminal threat, in contrast with the law-abiding majority, plays into the framing of protest as crime. According to Stevens, certain minority groups are specifically demonised and criminalised as a threatening 'other', in contrast to the 'general public'. Governments

can then use what Stevens names 'totemic toughness': seeming tough on the outsiders who are putting the general public and social order at risk (Stevens, 2011). In this way both the GRT community, and the Black community have been used to justify strengthening police powers and clamping down on protest.

The Gypsy Roma Traveller community's culture and way of living is often demonised by both the UK media (Tremlett, 2014) and UK government policy (Ryder and Cemlyn, 2016; for a full history of legislation against the GRT community, see Parnell-Berry and Lawton, 2018). In a study of the discourse on Travellers in the Houses of Parliament, Turner identified eight themes: 'criminal by nature', 'outside the community', 'menace', 'dirty', 'dishonest', 'immoral', 'nomadic' and 'real and fake'. The need for control was mentioned explicitly by several MPs from different parties, as a necessary means of dealing with Gypsy, Roma Travellers, who are characterised as 'criminal by nature' (Turner, 2002).

Scholars have also charted moral panics about the perceived increase and danger of youth crime (Pearson, 2014), which in the context of the 2011 Riots was also strongly racialised (Bridges, 2012). In an analysis of the findings of The Riots Panel, which was set up after the widespread civil unrest in August 2011, Bridges finds a focus on the defects of the rioters (such as absent fathers and lack of resilience or discipline) rather than any meaningful investigation of structural inequalities and societal issues (Bridges, 2012). The newspaper reporting echoed this framing of specifically Black youth as criminals (Clement, 2016), and recent government and media attention on the apparent threat of 'gang' activity and related knife crime also follows this trend (Hallsworth, 2013).

Relating hypotheses

H5 Minority groups will be associated with latent themes of criminality and the need for tighter restrictions on protest.

H6 The law-abiding majority will be associated with latent themes of protection from danger and threatening 'others'.

Data and Measures

Data

This study is interested in any debates taking place in the British House of Commons over the last 51 years (1970-2021) which relate specifically to political protest. In constructing the selection criteria for relevant debates, this study draws on previous research into extra-parliamentary activity. Prior studies have used the words 'protest,' 'demonstration', and 'strike' to conduct searches and reviews of 'extra-parliamentary activity' (Bailey, 2014; Morales, 2009), and considered activity to be political if the aims are considered to be to 'influence political decision-making processes' (Bailey, 2014). This search therefore included all debates specifically discussing any protest action or demonstration, significant strike action, or any debates around changes to legislation that affects the right to protest. Following previous studies, a decision was made to include the so-called 'Brixton Riots' of the 1980s and the 'UK riots' of 2011; given that both were born of an impetus to invoke political change (Bauman, 2011; Akram, 2014). The selected debates are shown in the table below.

Year	Number of debates	Protest/Legislation	Detail
1970	2	Industrial Relations Bill	Put limitations on strikes and increased the power of the courts in settling disputes.
1979	3	Winter of Discontent	Public and private sector strikes against government salary policy during high levels of inflation (Martin, 2009).
1981	2	Brixton Riots	Uprising in Brixton where anger about racist treatment of locals bubbled over into confrontation with the police (Peplow, 2019).
1984	2	Miners' Strikes	Coal industry strikes against government "Plan for Coal" policy (Beckett, 2009).
1985-6	2	Public Order Bill	Change to legislation, with implications on protest.
1990	1	Poll Tax Riots	Protest against the government's "Community Charge" (Bagguely, 1995).
1994	1	Criminal Justice and Public Order Bill	Proposed change to legislation, with implications on protest.
2009	2	G20 Protests	A member of the public was killed by police during protests about economic policy, the war on terror and climate change (The Guardian, 2009).
2010	2	NUS Tuition Fee Protests	Student protest against the rise in tuition fees.
2011	1	Anti-Austerity March	Protest against the Coalition's austerity measures.
2011	3	UK Riots	Uprising across the country following the murder of Mark Duggan by police in London (Briggs, 2012).
2020	1	Black Lives Matter	Protests against murder of Black people at the hands of the police, sparked by the murder of George Floyd in the U.S. (The Guardian, 2020)
2020	1	Extinction Rebellion	Protests to campaign for action on climate change.
2021	3	Police, Crime, Sentencing and Courts Bill	Change to legislation, with implications on protest.

Table 1: List of chosen debates for the analysis, with brief historic detail

Three other large protest marches were identified within the time period (1983 CND March for Peace, 2002 Liberty and Livelihood, and 2003 Stop the War), but were not discussed in a specific debate in the House of Commons, so were discounted.

The raw data were taken from Hansard, the government's online record of parliamentary debates (Hansard, 2021). Selected debates were downloaded in plain text format. Hansard formatting has varied over the time period selected, so some preparatory, manual cleaning of the data was done before reading the files into R. This consisted of adding MPs' constituency and party after their name where it was missing, which was only the case for older debates. The final raw data consisted of 28 debates, made up of 609,803 words.

Covariates

A data frame was constructed where each row of debate data was associated with metadata; including the speaker, the speaker's political party, the debate, and the year. The two covariates used in the analysis were the political party, and the year, to test hypotheses about the impact of political affiliation on protest discourse, and how this discourse has changed over time. The analysis focused on the two largest parties in the UK, the Conservative and Labour Party.

Analytical Strategies

Topic Modelling as Method

The development of machine-assisted textual analysis has enabled much larger amounts of texts to be analysed, which would often be impossible and/or expensive to analyse manually (Lucas et al., 2015). This expands the scope of text-based analysis, particularly with political texts, as government transcripts and records are being digitised in huge volume (Grimmer and Stewart, 2013). These relatively new methods have their issues: some advocates for qualitative methods warn that linguistic nuances, subtleties of language choice, tone and cultural references are impossible for an automated machine to comprehend (Ignatow and Mihalcea, 2017). As Grimmer and Stewart advise, automated textual analysis will never be able to fully capture the complexities of language; but it is a powerful supplementary tool for the human researcher (Grimmer and Stewart, 2013). Indeed, some scholars

perceive it to be underused in the social sciences, and encourage much more widespread use (Valdez et al., 2018).

Topic modelling includes several unsupervised methods for working with large corpora of text-based documents to reveal latent patterns and clusters of themes, which it does by identifying the co-occurrence of words across documents (Blei, 2012). It uses a 'bag of words' approach; analysing the interdependence of words rather than just each word independently. Latent Dirichlet Allocation (LDA) is one such statistical model, which finds latent topics across a large number of text documents (Blei, 2003). To investigate the hypotheses presented in this paper, the model needed to incorporate metadata as covariates, which is not possible with the LDA model (Blei, 2012). This analysis therefore uses the Structural Topic Model via the **stm** package. The Structural Topic Model built on the foundations of LDA, allowing for the user to analyse relationships between variables (Roberts et al., 2015). Structural topic modelling allows for metadata to be entered via topical prevalence (how covariates affect the frequency of topics appearing) and topical content (how covariates affect the words used within a given topic) (Roberts et al., 2019). The structural topic model is also mixed-membership, which allows for tokens to belong to multiple topics to different degrees (Blei, 2012; DiMaggio, 2013).

Pre-processing

Prior to fitting the model using **stm**, the analysis followed several pre-processing steps using **quanteda** (Benoit et al., 2018) and **tidytext** packages (Silge et al., 2018). The data were tokenised and all punctuation and capitalisation were removed (Jurafsky and Martin, 2009). The data were then stemmed and lemmatised, which reduces words down to the most common form of the word (e.g. *played* becomes *play*, and *went* becomes *go*) (Manning et al., 2008). Standard English stopwords and a set of user-specified stop words were also removed (a full list of these user-defined stop words can be seen in the code provided in the appendices). This reduced the text to 591,289 words of which 306,879 were from Conservative speeches and 238,550 from Labour speeches. Several descriptive analyses were carried out, to get an initial picture of the corpus being studied, and this informed the refinement of the stopwords and any necessary further cleaning of the data. The data were then converted to an **stm** format to feed into the structural topic model. A training dataset, made up of a random selection of parliamentary debates from the House of Commons, was used to test the model prior to entering the selected data for analysis (Hand, 2006).

Selecting the Models

When running automated textual analysis, we need to select a model which has the best substantive fit (Grimmer and Sewart, 2013). The **stm** package requires the user to select the number of topics the model will work with (K). There is not a set number of models to run, nor is there a set number of topics deemed appropriate for a given corpus size (Grimmer and Stewart, 2013). For corpora of several thousand words, authors of the **stm** package advise between 5-50 topics (Roberts et al., 2019). Using the searchK function, held-out likelihood, residuals, semantic coherence and exclusivity were tested to decide on the most preferable number of topics to feed into the **stm** processor (Roberts et al., 2016). Given the results in the graphs below, a model between 10 -20 topics was shown to be preferable.

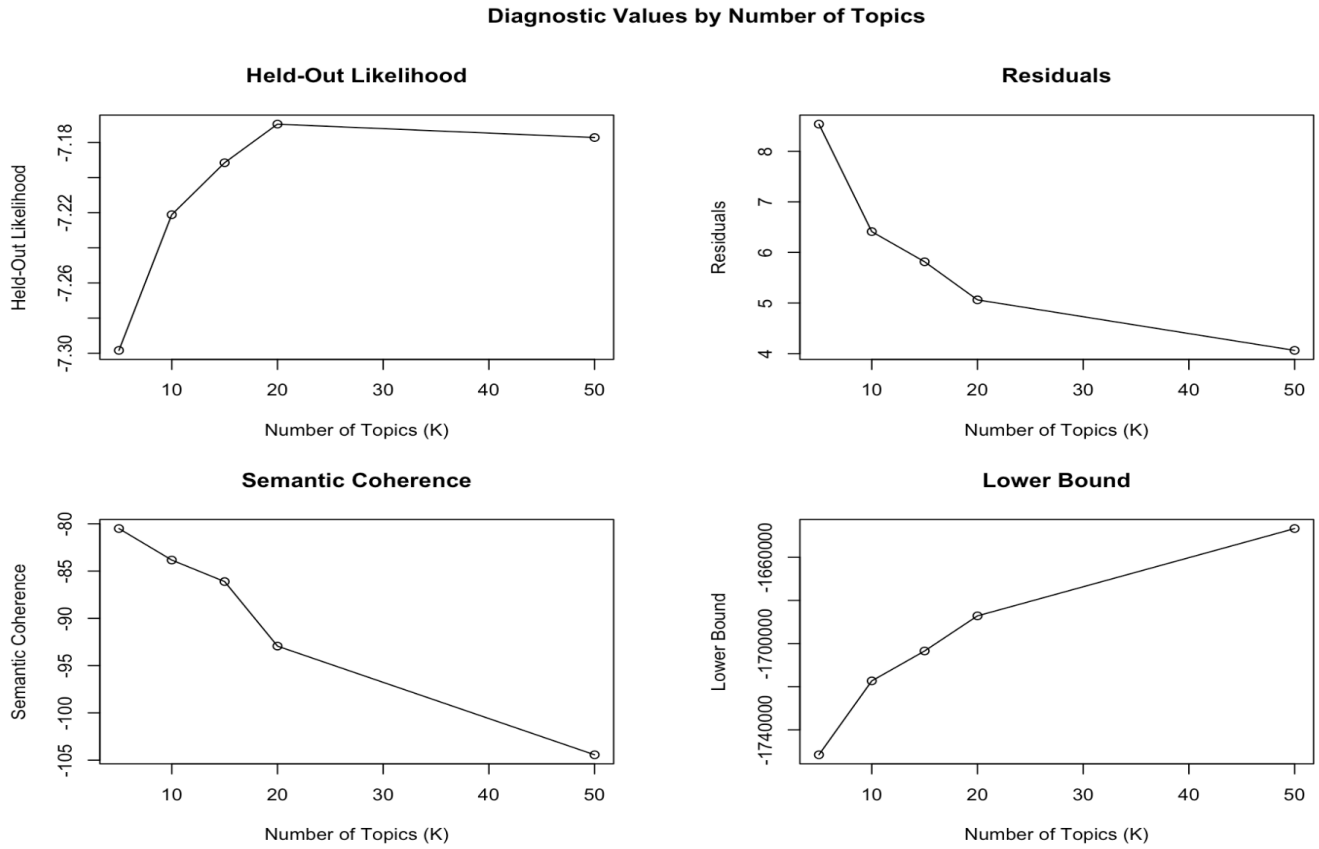

Figure 1: Results of diagnostic tests on models where K=10, 15 or 20

To select the exact number of topics, four topic models were modelled, with 12, 15, 17 and 20 topics. The exclusivity and semantic coherence for each was plotted, to find a model where semantic coherence and exclusivity to the given topic were both as strong as possible (Taddy, 2012). The four models tested were fairly similar, but the model with 12 topics had the least significant outliers, and so was chosen to bring forward in the analysis.

In running the topic model, Spectral initialisation was used, which selects an anchor word for each topic and then reconstructs the topic distributions based on the selected anchor words. This

serves to stabilise the overall model by using non-negative matrix factorisation, making the model easier to replicate (Arora et al., 2013).

Topical Prevalence and Topical Content Covariates

The stm package allows for covariates to be brought into the model in different ways: either as topical prevalence or topical content covariates (Roberts et al., 2019). Firstly, the political party was entered into the model as a topical prevalence covariate, to explore how prevalent the different topics were in speeches by the different political parties. The year was also entered additively, as a continuous covariate estimated with a spline (Roberts et al., 2019). This allowed for analysis of how topic prevalence varied over time.

Various different tools were then used to interpret the model, including word clouds weighted by highest word probability within each topic; quotations drawn from the corpus which were highly associated with specific topics, and FREX, lift and score words. FREX words are calculated by weighting words by both their overall frequency across the corpus, and how exclusive they are to a particular topic. This allows for common words that are highly correlated with a topic to show in their representative words (Roberts et al., 2015). Lift and score words weight by their low frequency in other topics, thereby rising up low frequency words that are strongly associated with a given topic (Taddy, 2013). The topics were assigned labels, based on the outputs and the literature outlined above.

A second model was estimated, using party and year as topical content covariates. The topics were interpreted using the same tools as the first model, and topics in both models that appeared to be referring to the same or similar latent themes were analysed alongside each other. Entering the party as a topical content covariate allowed the words associated with that topic to vary depending on the political party (Roberts et al., 2019). This enabled analysis of the different vocabulary used by each party when discussing the same topic.

Results

Overview of the Topics

Once the topic model had been selected, the topics were summarised with the associated highest probability words shown in the graphic below. Words were associated with topics based on the beta probability of belonging to each topic, and co-occurrence across topics was allowed (as can be seen with the stem 'polic', which appears in multiple topics to different degrees). A label was chosen to characterise the latent themes behind each topic, based on examining the associated words and most strongly associated speeches, and considering the existing literature.

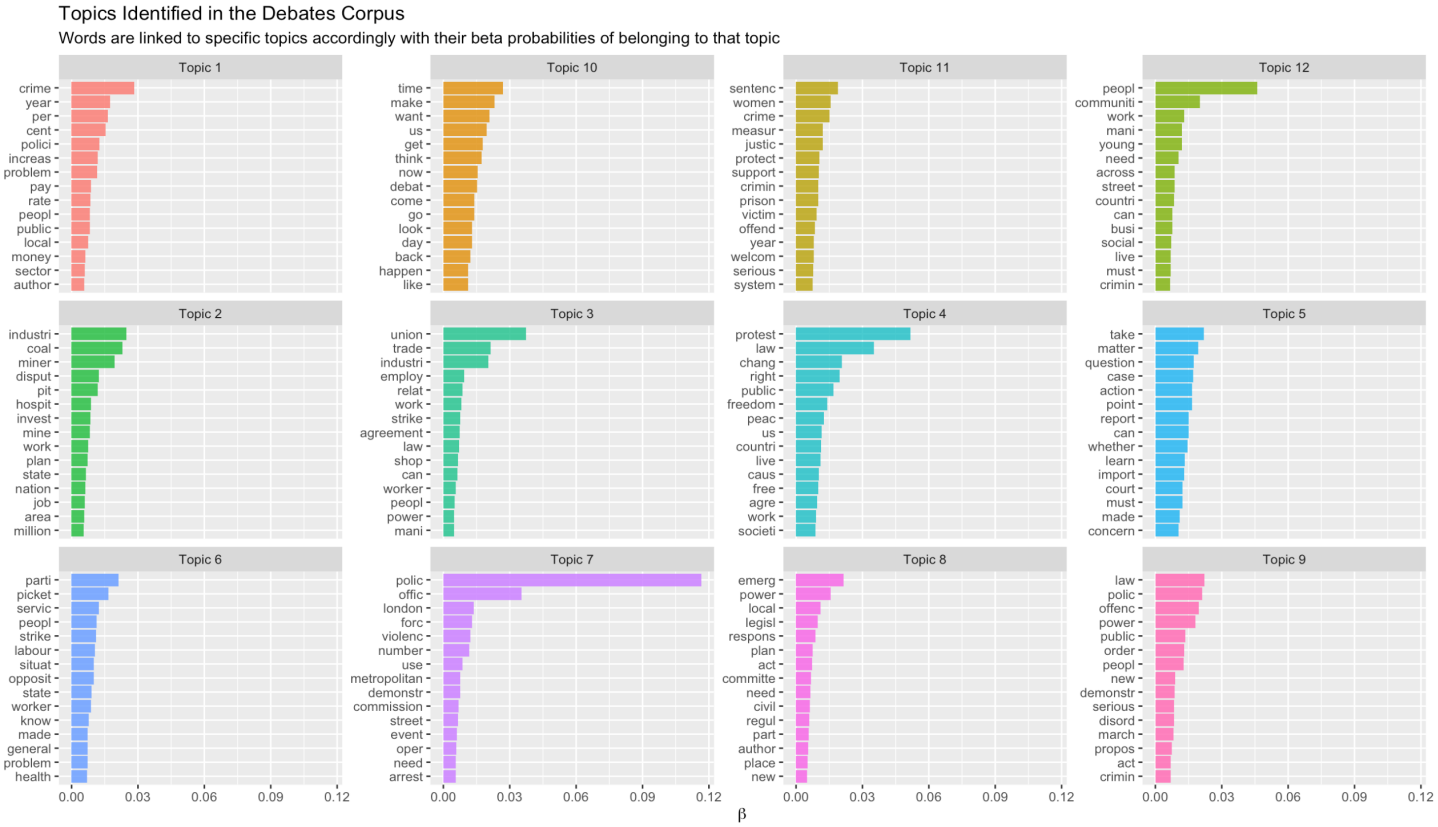

Table 2 shows the chosen topic labels, along with sections of the debates found to be most strongly associated with that topic. This gives an idea of the kind of speeches that were generating the latent patterns, and exploring this also supported in generating the chosen labels. It is immediately clear that the forms of protest have begun to be defined by different topics: topics 2, 3, and 6 relate to strike action and pickets, which is more strongly associated with the 1970s and 1980s. Topics 5 and 10 contained general words pertaining to debates, so were less relevant to this study and were dropped from the analysis.

Table 2: Topics with assigned labels and associated quotes
Increased crime and unemployment (Topic 1):
"There is a policy of high unemployment. Twentythree per cent of 16 to 19 year olds are jobless. In Barton Hill in my constituency, there is male unemployment of 45 per cent. There is a chronic housing shortage."
Coal dispute (Topic 2):
"No, I must not. There will be no victory in this dispute. The miners will become a less effective work force. They are damaging the future of their industry. As a result of this dispute, there will be uncertainty about the supply of coal. The memory of the dispute will remain with people for years, in company after company and industry after industry."
Trade unions (Topic 3):
"I agree only to the extent that where the parties, the trade union and the employer, have come to an agreement, there is created by the Bill a presumption of an intention between the parties that the contract shall be legally binding and enforceable."
Freedom to protest (Topic 4):
"Protest is the foundation of our democracy. Like many Members of Parliament, I have protested outside of this place for far longer than I have been within it. The right to protest must be protected for us all, and I will use my position here to do all that I can to defend it."
Pickets and strikes (Topic 6):
"Priority supplies, I am told, are moving out of a number of ports, although usually on a restricted basis and not at all at Hull or Felixstowe. Picketing of some food manufacturers and suppliers of animal feeding stuffs continues, but has eased."
Violence and policing protest (Topic 7):
"I want to be absolutely clear that the blame for the violence lies squarely and solely with those who carried it out. The idea advanced by some that police tactics were to blame, when people came armed with sticks, flares, fireworks, stones and snooker balls, is as ridiculous as it is unfair."
The need for emergency powers (Topic 8):
"The problem with the existing Acts is that they were drafted for a different time and different circumstances. The 1948 Act only requires local authorities to plan in preparation for hostile attack by a foreign power. The range of activities local authorities plan for far exceeds those set out in the Act and is therefore not adequately covered by legislation."
Disorder and power (Topic 9):
"I shall not repeat the way in which it described vividly the sort of disruption that can be caused by marches. We believe that the police should have the power to reroute a march to limit the resulting congestion of traffic, to prevent a bridge being blocked, for example, or to stop a city centre being brought to a standstill."
Victims and criminals (Topic 11):
"My constituents have waited a long time for the justice system to feel like it is putting victims before criminals, and this Bill will deliver that, with tougher sentences for assaulting emergency workers, stricter conditions on bail in high harm cases, including domestic abuse, increased jail time for sex offenders and child abusers, and extra funding for violence reduction, including knife crime. This Government are making our communities safer."
Riots and criminal acts (Topic 12):
"On Sunday morning, I stood amid the burning embers of Tottenham High road. There is no connection between the death of a young man and the torching of the homes of Stuart Radose and 25 other families in the Carpetright building. There is no connection between the treatment of the Duggan family and Niche, the landlord of the Spirit of Tottenham, being held at knifepoint while his pub was ransacked. I could go on. This violence was criminal, and we condemn it utterly."

In order to evaluate hypotheses relating to the difference between political parties and discussion of protest in the House of Commons, party was brought in as a covariate and the estimated topic proportion and confidence intervals were plotted by party (see figure 3).

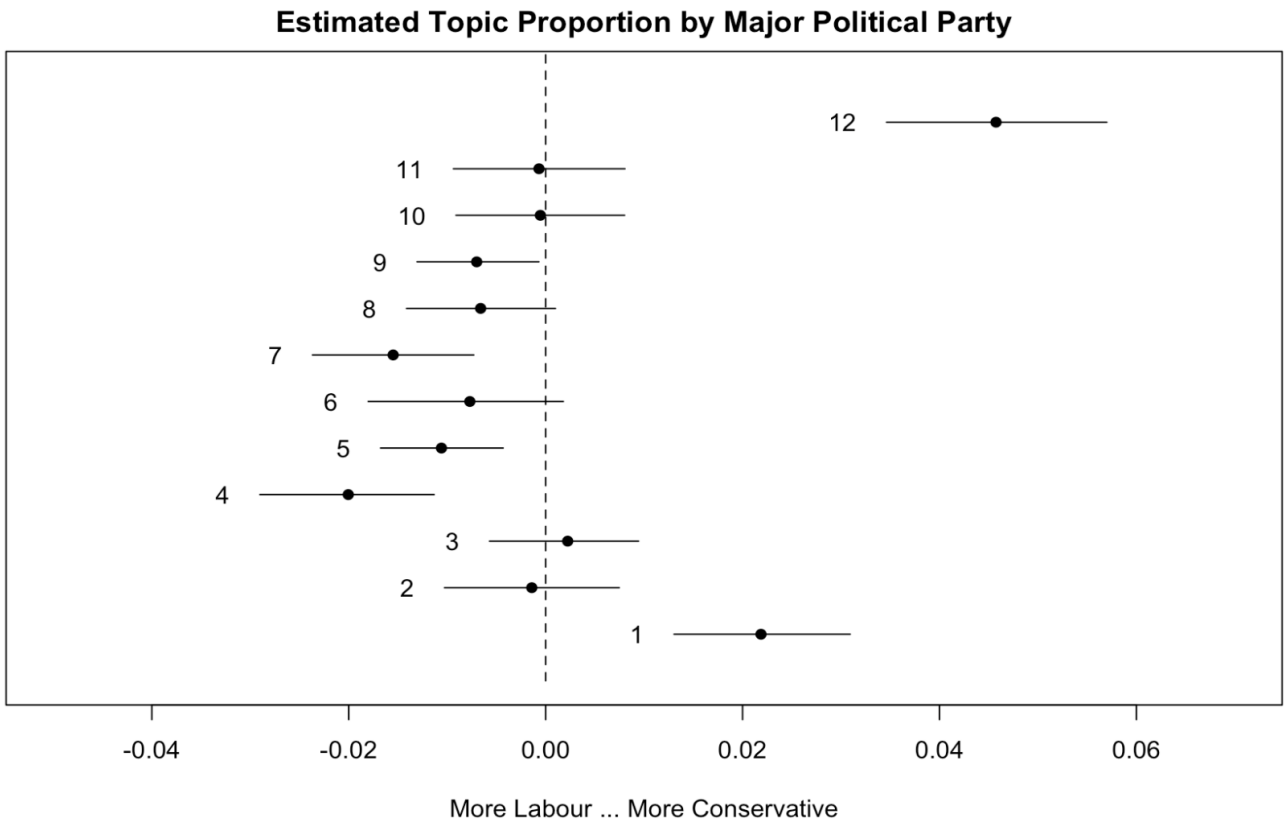

Figure 2: Estimated topic proportion to be discussed by political party (Labour vs Conservative)

The prevalence of topics associated with each party show some support for hypotheses 2 and 3. Topic 12 (riots and criminal acts) and topic 1 (increased crime and unemployment), which infer a preoccupation with crime and disorder, were much more prevalent in Conservative speeches compared with Labour. Topic 4 (freedom to protest) was most strongly associated with Labour, which reflects hypothesis 4 in showing a tendency for Labour MPs to frame protest in terms of democratic freedoms. However, topic 7 (police, violence and demonstrations) was also highly aligned with Labour speeches, which suggests a slightly different emphasis.

The latent topics will now be explored in more detail. Firstly, the analysis will consider how themes of power and social order play a part in debates about managing protest, testing hypothesis 1. Secondly, the discussion will turn to how protest action is framed by Labour and Conservative MPs, testing hypotheses 2, 3 and 4. Finally, the question of demonisation of protesters, and in particular of minority groups, will test hypotheses 5 and 6.

Power and Social Order

The topic classified as 'disorder and power' (topic 9) conceptually shows a clear concern for social order: police power and the rule of law are aligned with references to 'criminality', 'threat', 'disorder', 'disruption', and 'violence'. This echoes Maroto et al.'s findings, in presenting protest as a criminal threat to public order (Maroto et al., 2019). Within this topic we see references to marches, demonstrations and pickets: where these are mentioned there is an emphasis on disorder as a justification for the need for stronger powers, as is evident in the quote in Table 2. 'Control', 'stop' and 'ban' are actions associated with this topic, as well as 'protect', which works in opposition to 'threaten.' Topic 9 is heavily associated with the mid 1970s (estimated topic proportion = 4), suggesting that this latent pattern of converging disorder and police/legal power has shifted in more recent years.

Figure 3: Word cloud showing highest probability words associated with Topic 9, 'disorder and power'

Topic 6 in the topical content model has many crossover words with topic 9: its marginal highest probability words include 'law', 'public', 'order', 'power', 'disord[er]', 'march', 'demonstr[ation]', and 'peopl[e]'. We can surmise that the two topics represent the same latent theme, of using alleged threat to social order to justify state power. When the topic associated vocabulary is allowed to vary by political party, the strongest words within the topic to be used by the Conservative party include 'law', 'public', 'serious' and 'intimid[ation]'; words which suggest a focus on stronger laws to deal with the serious disorder which marches and demonstrations pose to the public. Labour speeches in this topic

seem to tend to be more about pickets and striking deals with trade unions, with high probability words like 'picket', 'deal' and 'propos[ition].

This supports hypothesis 1, in showing politicians framing protest as a threat to social harmony, and making the case for stronger restrictions on protest by 'fabricating the social order' (Neocleous, 2000). The stronger emphasis on these themes seen in the Conservative speeches also shows support for hypothesis 2.

Framing Protest

To turn to the explicit framing of protest marches and demonstrations, topic 4 stands out as relating clearly to the legal freedom to protest. The highest probability words include words relating to the legal right to protest, such as 'freedom', 'right', 'fundament[al]', and 'citizen'. The topic also contains methods of protest: 'peace[ful]', 'annoy', 'disrupt' and 'illegal'. Unlike several of the other topics, this topic is prevalent fairly consistently from 1980 onwards, with peaks around 2010 and 2020, which fits the 'cycle of contentions' said to peak at around the same times (Clement, 2019; Gerbaudo, 2020; Chessum, 2021). Topic 4 is strongly skewed towards Labour in Figure 3, suggesting that this framing of protest as a fundamental freedom and right is employed more consistently by Labour MPs as compared with Conservatives. Figure 5 shows the different vocabulary associated with this topic in Labour speeches compared with Conservative.

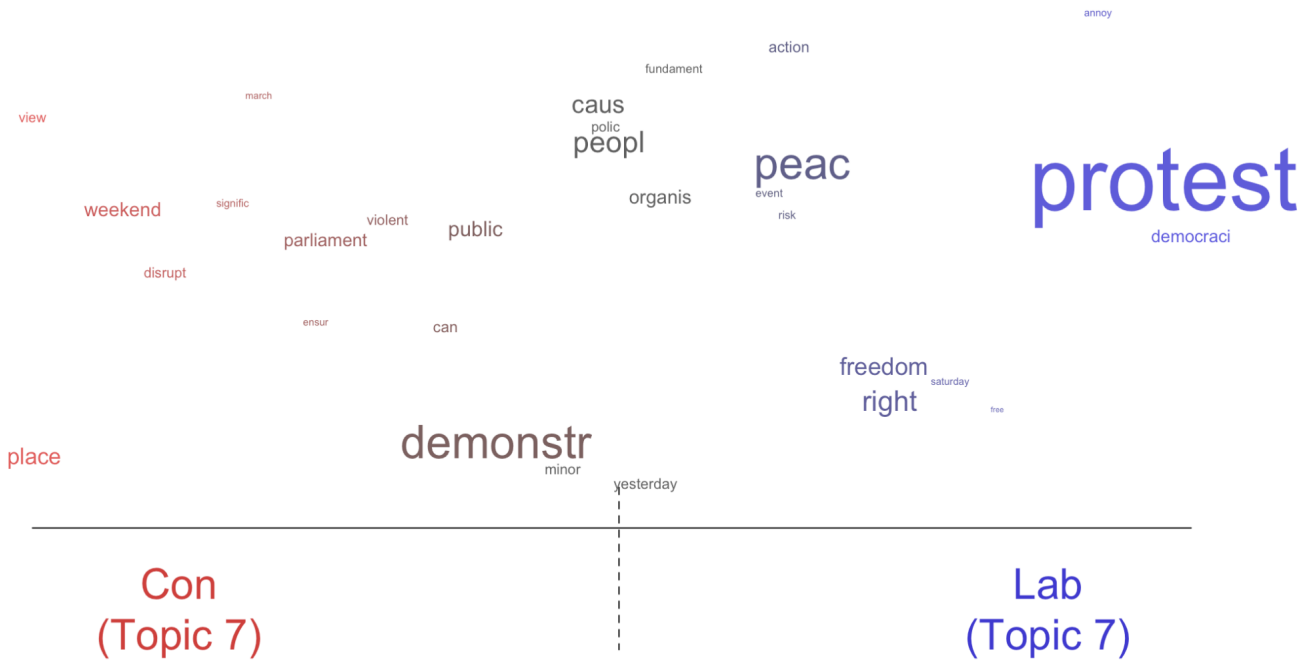

Figure 4: Graphical display showing vocabulary used by Labour and Conservative MPs when discussing the latent theme, 'freedom to protest'

Hypothesis 4 predicted that protest would be framed in terms of criminality, rather than in terms of politics. Conservatives are more likely to use the word 'demonstr[ation]' within this topic, whereas with Labour the word 'protest' is much more strongly associated. Protest is a more politically charged word, evoking anger, action, and dissent. Demonstration, on the other hand, feels more organised, controlled, and depoliticised. Alongside the other frequent words for the Conservative party, 'weekend', 'place' and 'parliament', it seems that speeches associated with this topic are alluding to specific events that have taken place; this was also shown to be the case when the associated quotes for this topic were reviewed. The words 'violent', 'disrupt' and 'public' also suggest a criminal framing. Labour's commonly associated words 'peac[eful]', 'right', 'freedom', and 'democrac[y]' allude more towards the overarching right to protest, and protest's place in democracy.

To focus in on the Conservative framing of protest, the analysis looked at the commonly co-occuring words with 'demonstration' and 'protester', based on their phi coefficients. When looking at words commonly co-occuring with 'protesters' in Conservative MP speeches, 'peaceful' comes second to 'arrests', and common words also include 'thugs' and 'tactics', suggesting a focus on the criminality of protesters. Interestingly, the word 'demonstration' also commonly co-occurs with 'damage' and 'property', which supports the theory that policing is about governments fabricating social order and protecting state property and institutions (Waddington, 1987; Neocleous, 2000).

Demonisation and the Law-Abiding Majority

One of the most disturbing latent themes to emerge from the analysis, is topic 11, which at first glance seems to relate to crimes against women, sexual assault, and associated criminal justice and legal issues. The highest probability words include 'sentence', 'women', 'crime', 'justice', 'protect' and 'support.' We might assume that this topic is highly associated with the year 2021, as many of the words seem likely to be drawn from debates discussing the Sarah Everard vigil. However, the prevalence of this topic also rises around 1994 and in the mid 2010s.

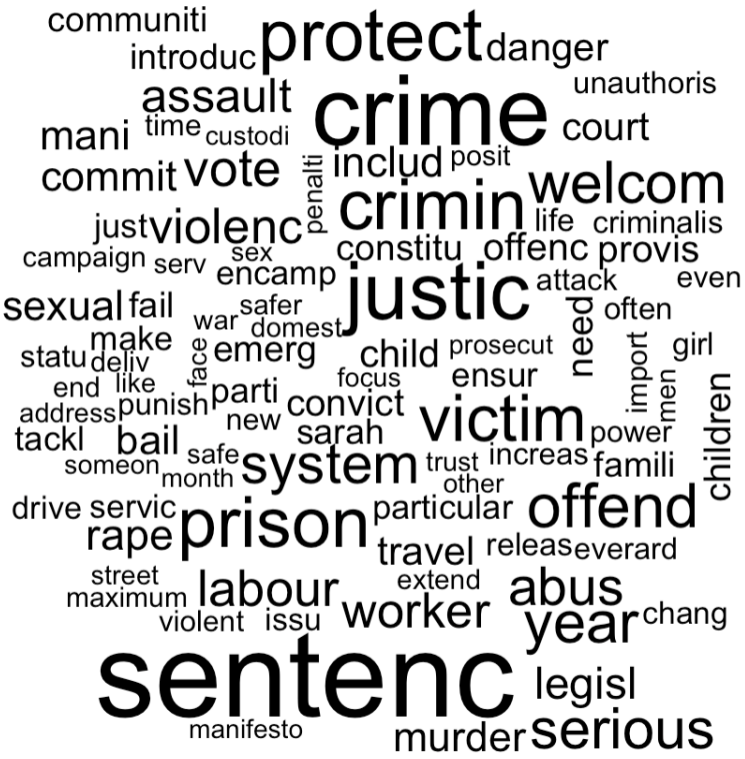

Figure 5: Word cloud showing highest probability words associated with Topic 11, 'victims and criminals'

When the FREX words as well as the highest probability words are considered, another aspect to this topic emerges. The top FREX words are 'sarah, encamp[ment], unauthoris[ed], everard, sex, gyps[y], coven[tion]'. When weighted for their overall frequency and exclusivity to a particular topic, words associated with the Gypsy Roma Traveller community, such as 'gyps[y]' and 'encamp[ment]' are included. When we look at most probable bigrams with the words 'traveller', 'gypsy' or 'encampment', 'unauthorised' appears, so this is likely to also be an indicator that discussion of the GRT community is part of this latent theme (other bigrams include 'illegal' and 'so-called').

This supports hypothesis 5 in linking minority groups (here the GRT community) with latent themes of criminality. This topic also shows an underlying conceptual association between protecting women and the alleged criminality of minority groups, thereby supporting hypothesis 6. Using alleged threat towards [white] women as a pretext for racism is an age-old trope (McMahon and Kahn, 2018), and one that appears as a latent pattern in these debates. This finding goes far beyond the expected distinction between the law-abiding public under threat, and the criminalised minority.

Conclusion

This article began by referencing the many times in recent UK history that governments have introduced new legislation, each time more restrictive, attempting to set out where the limits are when it comes to political protest. Protest is, by its very nature, disruptive; and the findings here reveal the challenge that such disruption causes to governments attempting to 'fabricate a social order' (Neocleous, 2000). Staying in control and maintaining power means framing protest in a specific way in the House of Commons. The findings of this analysis reveal latent themes bringing together social order, police power, and the threat of disorder. By depoliticising protest, and aligning it strongly with criminality, governments are able to justify the need for strong police power.

The analysis suggests that there are still distinct differences in the way that Labour and Conservative MPs discuss protest, despite previous research suggesting that Labour has weakened its connection with the trade unions and working-class movements during the time period studied (Kogan, 2019). Labour speeches were more likely to use the more politicised word, 'protest', and were more likely to contain themes of protest as legal right; the importance of freedom to protest. Conservative speeches, on the other hand, were more associated with themes of legality, crime and threat to the general public. The word 'demonstration' was used much more frequently by Conservative MPs, and demonstrations and protesters were associated with property damage, disruption, violence and arrests.

This preoccupation with disruption and disorder is interesting in revealing what political protest represents to those in power. There is an inherent need for citizens to conform into the ordered system, and not going to work; blocking the roads; bringing noise and chaos to the streets, and damaging property threatens the order that policing is there to maintain (Waddington, 1987; Butler, 2006). That the Gypsy Roma Traveller community, a community whose way of life does not fit within this ordered system, are part of this discussion shows how inherent order is to these debates. Maintaining control through manufactured social order means those seen as deviant, whether that is the GRT community's nomadic way of life, or protesters' choice to engage in disruptive action, are criminalised both in terms of parliamentary discourse, but also through the restrictive legislation that follows.

When testing the effect of political party on the latent themes, this analysis was limited to comparing Conservative and Labour party speeches in any meaningful way. The Liberal Democrats were in coalition power for a period of time during this analysis, and exploring their association with the themes could shed further light on the research questions. Equally, exploring how being in power versus being in opposition affected association with the topics could reveal more about how holding power affects a party's relationship with political protest. Despite the limitations, this article reveals

some strong latent themes in parliamentary discussion of protest, and demonstrates the usefulness of structural topic modelling as a method to work with parliamentary transcripts. If we are on the verge of a new cycle of contention, with further protests and civil unrest on the horizon in the UK, future research using structural topic modelling could explore the role of the media in framing protest, and use social media scraping to evaluate how the public view upcoming protest movements.

References

Akram, S., 2014. Recognizing the 2011 United Kingdom Riots as Political Protest. *British Journal of Criminology*, 54(3), pp.375–392.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y. & Zhu, M.. 2013. A Practical Algorithm for Topic Modeling with Provable Guarantees. *Proceedings of the 30th International Conference on Machine Learning, in PMLR* 28(2):280-288

Bailey, D., 2014. Contending the crisis: What role for extra-parliamentary British politics? British Politics, 9(1), pp.68–92.

Baker, P., 2004. 'Unnatural Acts': Discourses of homosexuality within the House of Lords debates on gay male law reform. *Journal of Sociolinguistics*, 8(1), pp.88–106.

Bauman, Z. 2011. The London Riots – On consumerism coming home to roost, *Social Europe Journal*.

Benoit, Kenneth et al., 2018. Quanteda/Quanteda: (Cran) V1.3.10.

Blei, D., 2012. Probabilistic topic models. *Communications of the ACM*, 55(4), pp.77–84.

Blei, D., Jordan, M. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res. 3* (January 2003), 993–1022.

Butler, J., 2006. Precarious life : the powers of mourning and violence, London: Verso.

Chessum, M., 2021. In the UK, a long, hot summer of post-lockdown protests has begun. [online] openDemocracy. Available at: <https://www.opendemocracy.net/en/opendemocracyuk/long-hot-summer-post-lockdown-protests-has-begun/> [Accessed 9 May 2021].

Clark, N. and Davidson, L., 2017. Thousands of masked protesters throw smoke bombs and wave 'Tory scum' banners. [online] *The Sun*. Available at: <https://www.thesun.co.uk/news/4589263/conservative-party-conference-protest-anti-tory-march-pictures/> [Accessed 10 May 2021].

Clement, M., 2019. Race for the Future. Social Justice, 46(2/3), pp.125–142.

Clement, Matt., 2016. *The Sound of the Crowd: A People's History of Riots, Protest and the Law*. London: Palgrave Macmillan.

della Porta, Donatella., 2008. "Eventful Protest, Global Conflicts." *Journal of Social Theory* 9(2): 27–56.

DiMaggio, P., Nag, M., & Blei, D. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics*, 41(6), 570–606.

Extinction Rebellion, 2021. *#KillTheBill: Joint Statement on the Police, Crime, Sentencing and Courts Bill From XR, BLM local groups, RAAH and more*. Available at: <https://extinctionrebellion.uk/2021/03/15/killthebill-joint-statement-on-the-police-crime-sentencing-and-courts-bill-from-xr-blm-local-groups-raah-and-more//> [Accessed 12 May 2021].

Fernandez, L.A., 2009. *Policing dissent : social control and the anti-globalization movement*, New Brunswick, N.J.: Rutgers University Press.

Ford, M. & Novitz, T., 2015. An Absence of Fairness... Restrictions on Industrial Action and Protest in the Trade Union Bill 2015. *The Industrial Law Journal*, 44(4), p.522.

Garland, D., 1991. Sociological Perspectives on Punishment. *Crime and Justice*, 14, pp.115–165.

Garland, D., 2001. *The Culture of Control – crime and social order in contemporary society*, Oxford: Oxford University Press.

Gerbaudo, P., 2020. The Pandemic Crowd: protest in the time of COVID-19. *Journal of International Affairs*, 73(2), pp.61–75.

Grimmer, J. & Stewart, B.M., 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), pp.267–297.

Hallsworth, S., 2013. *The Gang and Beyond : Interpreting Violent Street Worlds,* Houndmills, Basingstoke.

Hand, David J. 2006. Classifier technology and the illusion of progress. *Statistical Science* 21(1):1–15.

Hausser J, Strimmer K (2009). "Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks." *Journal of Machine Learning Research*, 10(Jul), 1469–1484.

Ignatow, G. And Mihalcea, R., 2017. *Text Mining: A Guidebook for the Social Sciences,* Los Angeles: SAGE.

Jenkins, J.C., Wallace, M. & Fullerton, A.S., 2008. A Social Movement Society?: A Cross-National Analysis of Protest Potential. *International Journal of Sociology*, 38(3), pp.12–35.

Jones, S., Percival, J. and Lewis, P., 2009. *G20 protests: riot police clash with demonstrators.* [online] The Guardian. Available at: <https://www.theguardian.com/world/2009/apr/01/g20-summit-protests> [Accessed 12 May 2021].

Jurafsky, Dan, and James Martin. 2009. *Speech and natural language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Kent, T., Cooke, A. & Marsh, I., 2020. "The expert and the patient": a discourse analysis of the house of commons' debates regarding the 2007 Mental Health Act. *Journal of Mental Health*, pp.1–6.

Kettell, S. & Kerr, P., 2021. The Brexit Religion and the Holy Grail of the NHS. *Social Policy and Society*, 2021(2), pp.282–295.

Kirkwood, S., 2017. The Humanisation of Refugees: A Discourse Analysis of UK Parliamentary Debates on the European Refugee 'Crisis.' *Journal of Community & Applied Social Psychology*, 27(2), pp.115–125.

Kogan, D., 2019. *Protest and Power*. London: Bloomsbury Publishing Plc.

Lee, M. & Mimno, D., 2017. Low-dimensional Embeddings for Interpretable Anchor-based Topic Inference.

Lee, M. and McGovern, A., 2013. Force to sell: Policing the image and manufacturing public confidence. *Policing and Society* 23: 103–124.

Lucas, C. et al., 2015. Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, 23(2), pp.254–277.

Ludlam, S., 2000. 'Norms and Blocks: Trade Unions and the Labour Party since 1964', in *The Labour Party: A Centenary History*, ed. by Brian Brivati and Richard Heffernan. Basingstoke: Macmillan

Ludovic Rheault et al., 2016. Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. *PloS one*, 11(12).

Manning, C. D., P. Raghavan, and H. Schutze. 2008. *Introduction to information retrieval*, Vol. 1. Cambridge: Cambridge University Press.

Margaret E. Roberts, Brandon M. Stewart & Dustin Tingley, 2019. stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, 91(1), pp.1–40.

Maroto, M., González-Sánchez, I. & Brandariz, J., 2019. Editors' Introduction: Policing the Protest Cycle of the 2010s. *Social Justice*, 46(2/3), pp.1–27.

Mawby, R.C., 2014. The presentation of police in everyday life: Police–press relations, impression management and the Leveson Inquiry. *Crime, media, culture*, 10(3), pp.239–257.

McClelland, J., 1989. *The Crowd and the Mob*. London: Unwin Hyman.

McIlroy, John. 1998. 'The Enduring Alliance? Trade Unions and the Making of New Labour, 1994−1997', *British Journal of Industrial Relations*, 36.4, 537−64.

McMahon, J.M. and Kahn, K.B., 2018. When sexism leads to racism: Threat, protecting women, and racial bias. *Sex Roles*, 78(9), pp.591-605.

Mead, D., 2021. *The Police, Crime, Sentencing and Courts Bill reinforces tensions and division at the expense of collective social solidarity*, [online] LSE Blogs. Available at: <https://blogs.lse.ac.uk/politicsandpolicy/police-crime-sentencing-courts-bill/> [Accessed 12 May 2021].

Minkin, Lewis. 1991. *The Contentious Alliance: Trade Unions and the Labour Party,* Edinburgh: Edinburgh University Press.

Mohammad, S.M. et al., 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4), pp.480–499.

Morales, L. 2009. *Joining Political Organisations: Institutions, Mobilisation and Participation in Western Democracies*. Essex, UK: ECPR Press.

Neocleous, M., 2000. *The fabrication of social order : a critical theory of police power*, London: Pluto Press.

Opp, K.-D.,2009. *Theories of Political Protest and Social Movements: A Multidisciplinary Introduction, Critique, and Synthesis*, London: Routledge.

Parnell-Berry, B. & Lawton, A., 2019. Citizenship for All? Mobility and the "Right" Way to Live. *Administrative theory & praxis*, 41(1), pp.35–59.

Peplow, S., 2019. *Race and riots in Thatcher's Britain*, Manchester: Manchester University Press.

Roberts, M., Stewart, B. and Tingley, D., 2016. Navigating the Local Modes of Big Data: the case of topic models. In: R. Alvarez, ed., *Computational social science: Discovery and Prediction*. Cambridge: Cambridge University Press.

Roberts, Margaret, Stewart, B., & Airoldi, E., 2016. A model of text for experimentation in the social sciences. *Harvard Dataverse*.

Ryder, A. & Cemlyn, S., 2016. Monoculturalism, austerity and moral panics: assessing government progress on addressing Gypsy, Traveller and Roma exclusion. The Journal of Poverty and Social Justice, 24(2), pp.143–155.

Schweingruber, David. 2000. "Mob Sociology and Escalated Force: Sociology's Contribution in Repressive Police Tactics." *The Sociological Quarterly* 41 (3): 374-389.

Silge, D., Robinson, J. & Robinson, David, 2017. Text mining with R a tidy approach First.

Silge, Julia et al., 2018. Juliasilge/Tidytext: Tidytext 0.2.0.

Sombatpoonsiri, J., 2015. Articulation of Legitimacy: A Theoretical Note on Confrontational and Nonconfrontational Approaches to Protest Policing. *Asian Journal of Peacebuilding*, 3(1), pp.1–17.

Stevens, A. 2011. 'Telling policy stories: An ethnographic study of the use of evidence in policy-making in the UK. *Journal of Social Policy*, 40(2):237–255.

Stewart, D., 2011. Preserving the 'Contentious Alliance'? The Labour Party, the Trade Unions, and the Political Fund Ballots of 1985-1986. *Labour History Review*, 76(1), pp.51–69.

Taddy, M.A., 2012. "On Estimation and Selection for Topic Models." In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistic*s.

Tarrow, S. (1998) *Power in Movement: Social Movements and Contentious Politics*. 2nd edn. Cambridge: Cambridge University Press (Cambridge Studies in Comparative Politics).

The Guardian. 2020. *Hundreds join march to protest against systemic racism in the UK*. [online] Available at: <https://www.theguardian.com/world/2020/aug/30/hundreds-join-march-to-protest-against-systemic-racism-in-the-uk> [Accessed 12 May 2021].

Tremlett, A., 2014. Demotic or Demonic? Race, Class and Gender in 'Gypsy' Reality TV. The Sociological review (Keele), 62(2), pp.316–334.

Turner, R., 2002. Gypsies and British parliamentary language: An analysis. *Romani Studies*, 12(1), pp.1–34.

UK Government, 2021. Police, Crime, Sentencing and Courts Bill (as Introduced). House of Commons.

UN General Assembly, 1948. *Universal Declaration of Human Rights*, 10 December 1948, 217 A (III), available at: https://www.refworld.org/docid/3ae6b3712c.html [accessed 12 May 2021]

Valdez, . 2018.

Waddington, P.A.J. 1987. "Towards Paramilitarism? Dilemmas in Policing Civil Disorder." *British Journal of Criminology* 27 (1): 37-46.

Debates

HC Deb 14 December 1970, vol 808
HC Deb 15 December 1970, vol 808
HC Deb 16 January 1979, vol 960
HC Deb 25 January 1979, vol 961
HC Deb 6 February 1979, vol 962
HC Deb 13 April 1981, vol 3
HC Deb 25 November 1981, vol 13
HC Deb 7 June 1984, vol 62
HC Deb 19 June 1984, vol 56
HC Deb 16 May 1985, vol 60
HC Deb 13 January 1986, vol 89
HC Deb 7 November 1986, vol 235
HC Deb 2 April 1990, vol 101
HC Deb 11 January 1994, vol 234
HC Deb 12 May 2009, vol 72
HC Deb 18 May 2009, vol 72
HC Deb 11 November 2010, vol 571
HC Deb 13 December 2010, vol 589
HC Deb 28 March 2011, vol 784
HC Deb 11 August 2011, vol 811
HC Deb 11 August 2011, vol 811
HC Deb 11 August 2011, vol 811
HC Deb 8 June 2020, vol 13
HC Deb 15 June 2020, vol 25
HC Deb 7 September 2020, vol 98
HC Deb 15 March 2021, vol 691
HC Deb 16 March 2021, vol 691
HC Deb 18 March 2021, vol 691

Code used for Analysis

Calling packages

```
library(dplyr)
library(tidyr)
library(purrr)
library(readr)
library(stringr)
library(tidyverse)
library(tidytext)
library(quanteda)
library(stm)
library(readtext)
library(furrr)
library(formattable)
library(reshape2)
library(tm)
library(ggplot2)
library(forcats)
library(igraph)
library(ggraph)
library(widyr)
library(ggrepel)
library(topicmodels)
library(corpustools)
library(tibble)
library(vctrs)
library(devtools)
library(Rtsne)
library(geometry)
library(rsvd)
library(stmCorrViz)
```

Reading in files and creating raw text tibble

```
# add debates to a folder, with sub-folders by year
debate_folder <- "debates"

# define a function to read all files from a folder into a data
frame
```

```
read_folder <- function(infolder) {
  tibble(file = dir(infolder, full.names = TRUE)) %>%
    mutate(text = map(file, read_lines)) %>%
    transmute(debate = basename(file), text) %>%
    unnest(text)
}

# use unnest() and map() to apply read_folder to each subfolder
raw_text <- tibble(folder = dir(debate_folder, full.names =
TRUE)) %>%
  mutate(folder_out = map(folder, read_folder)) %>%
  unnest(cols = c(folder_out)) %>%
  transmute(year = basename(folder), debate, text)
```

Extracting metadata from raw text and adding to data.frame; cleaning data.frame

```
# taking out hyphens from names
raw_text[] <- lapply(raw_text, gsub, pattern='-', replacement='')

# detecting name formats to splice by speech
prepped_text <- raw_text %>%
  mutate(linenumber = row_number(),
         speech = cumsum(str_detect(text,
                                    regex("^[M](.*?)[.]
[[:upper:]](.*?) [[:upper:]](.*?) [(](.*?)[)]
[(](.*?)[)]$|^[[:upper:]](.*?) [[:upper:]](.*?) [(](.*?)[)]
[(](.*?)[)]$"))))

# select by first line (speaker name), and then extract the party
from the name
speakers <- prepped_text %>%
  group_by(speech) %>%
  slice(1) %>%
  select(-linenumber) %>%
  rename(speaker = text) %>%
  mutate(speaker = str_remove(speaker, " \\[V\\]"),
         party = str_extract(speaker, "(?<=\\()[A-z]*(?=\\)$)"),
         plain_name = str_remove(speaker, "\\(.*\\)"))

# add the speaker and party to the dataframe
```

```
prepped_text <- prepped_text %>% left_join(speakers, by =
c("speech" = "speech"))

# tidying new prepped_text tibble
prepped_text <- prepped_text %>%
  select(-year.y, -debate.y)

# ensuring all columns are classed correctly
prepped_text$year.x <-
as.integer(as.character(prepped_text$year.x))

# replacing missing values
prepped_text <- prepped_text %>%
  mutate(party = replace(party, is.na(party), "none"))

# now that speaker and party have been extracted, taking them out
of the main text
prepped_text$text<-str_remove(prepped_text$text,
regex("^[M](.*?)[.] [[:upper:]](.*?) [[:upper:]](.*?) [(](.*?)[)]
[(](.*?)[)]$|^[[:upper:]](.*?) [[:upper:]](.*?) [(](.*?)[)]
[(](.*?)[)]$"))

# removing times
prepped_text$text<-str_remove(prepped_text$text,
regex("^[[:digit:]][[:digit:]][:][[:digit:]][[:digit:]][:][[:digi
t:]][[:digit:]]$|^[[:digit:]][.][[:digit:]][[:digit:]][
][[:lower:]][.][[:lower:]][.]$|^[[:digit:]][.][[:digit:]][[:digit
:]][
][[:lower:]][[:lower:]]$|^[[:digit:]][[:digit:]][.][[:digit:]][[:
digit:]][ ][[:lower:]][[:lower:]]$"))

# removing empty rows
prepped_text <- prepped_text[!(prepped_text$text == "" |
is.na(prepped_text$text)), ]
prepped_text <- prepped_text[!(prepped_text$text == " " |
is.na(prepped_text$text)), ]
prepped_text <- prepped_text[!(prepped_text$text == "  " |
is.na(prepped_text$text)), ]

# adding row ID
prepped_text <- tibble::rowid_to_column(prepped_text, "rowID")
```

Preparing for stm() input: tokenising and stopwords

```
# creating corpus from prepped_text including docvars
pro_corpus <- corpus(
  prepped_text,
  docid_field = "rowID",
  text_field = "text",
  meta = list("year.x", "debate.x", "linenumber", "speech",
"speaker", "party", "plain_name"),
  )

# preparing for stm piped into one command
stm_input <- pro_corpus %>%
  tokens(remove_numbers = TRUE, remove_punct = TRUE,
remove_symbols = TRUE, include_docvars = TRUE) %>%
  tokens_tolower() %>%
  tokens_remove(pattern = c(stopwords("english"), "con", "lab",
"one", "house", "hon", "honourable", "right", "gentleman",
"bill", "government", "bill", "mr", "member", "secretary",
"speaker", "home", "prime", "minister", "friend", "will", "say",
"said", "also", "may", "priti", "patel", "give", "way",
"amendment", "clause", "maiden", "speech", "ms", "constituency",
"members", "friend")) %>%
  tokens_wordstem() %>%
  dfm() %>%
  convert(to = "stm")

# setting documents, vocab and metadata for stm
docs <- stm_input$documents
vocab <- stm_input$vocab
meta <-stm_input$meta
```

Running diagnostics to decide on the number of topics

```
# run diagnostics to search for best number for K
storage1 <- searchK(docs,
                    vocab,
                    K = c(5,10,15,20,50),
                    prevalence =~ party + s(year.x),
                    data=meta,
```

```
                    set.seed(9999),
                    verbose=TRUE
                    )

# plot these results for inspection
print(storage1$results)
options(repr.plot.width=6, repr.plot.height=6)
plot(storage1)

# run models for 12, 15, 17 and 20 topics
model12<-stm(docs,
             vocab,
             prevalence =~ party + s(year.x),
             K=12,
             data=meta,
             init.type = "Spectral",
             verbose=TRUE
             )

model15<-stm(docs,
             vocab,
             prevalence =~ party + s(year.x),
             K=15,
             data=meta,
             init.type = "Spectral",
             verbose=TRUE
             )

model17<-stm(docs,
             vocab,
             prevalence =~ party + s(year.x),
             K=17,
             data=meta,
             init.type = "Spectral",
             verbose=TRUE
             )

model20<-stm(docs,
             vocab,
             prevalence =~ party + s(year.x),
             K=20,
             data=meta,
```

```
              init.type = "Spectral",
              verbose=TRUE
              )

# plot these 3 models to compare semantic coherence and
exclusivity and select best model
suppressWarnings(library(ggplot2))
suppressWarnings(library(plotly))

M12ExSem<-as.data.frame(cbind(c(1:12),
                              exclusivity(model12),
                              semanticCoherence(model=model12,
docs),
                              "Mod10")
                        )

M15ExSem<-as.data.frame(cbind(c(1:15),
                              exclusivity(model15),
                              semanticCoherence(model=model15,
docs),
                              "Mod15")
                        )

M17ExSem<-as.data.frame(cbind(c(1:17),
                              exclusivity(model17),
                              semanticCoherence(model=model17,
docs),
                              "Mod17")
                        )

M20ExSem<-as.data.frame(cbind(c(1:20),
                              exclusivity(model20),
                              semanticCoherence(model=model20,
docs),
                              "Mod20")
                        )

ModsExSem<-rbind(M12ExSem, M15ExSem, M17ExSem, M20ExSem)

colnames(ModsExSem)<-c("K","Exclusivity", "SemanticCoherence",
"Model")
```

```
ModsExSem$Exclusivity<-
as.numeric(as.character(ModsExSem$Exclusivity))
ModsExSem$SemanticCoherence<-
as.numeric(as.character(ModsExSem$SemanticCoherence))

options(repr.plot.width=7, repr.plot.height=7, repr.plot.res=100)

plotexcoer<-ggplot(ModsExSem, aes(SemanticCoherence, Exclusivity,
color = Model))+geom_point(size = 2, alpha = 0.7) +
geom_text(aes(label=K), nudge_x=.05, nudge_y=.05)+
  labs(x = "Semantic coherence",
       y = "Exclusivity",
       title = "Comparing exclusivity and semantic coherence")

plotexcoer
```

Plotting selected structural topic model

```
# plot the selected model
plot(
  model12,
  type = "summary",
  n = 15,
  text.cex = 0.5,
  main = "STM topic shares",
  xlab = "Share estimation"
)

# plot word clouds for each topic
par(mar=c(0.5, 0.5, 0.5, 0.5))
cloud(model12, topic = 7, scale = c(2.25,.5))

# or more simply
cloud(model12, topic = 7)
```

Plotting the topics using Tidyverse

```
# plot overall tidy graphic showing beta scores for each topic
td_beta <- tidytext::tidy(model12)

td_beta %>%
```

```
  group_by(topic) %>%
  top_n(15, beta) %>%
 ungroup() %>%
    mutate(topic = paste0("Topic ", topic),
           term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = as.factor(topic))) +
  geom_col(alpha = 0.8, show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free_y") +
  coord_flip() +
  scale_x_reordered() +
  labs(x = NULL, y = expression(beta),
       title = "Highest word probabilities for each topic",
       subtitle = "Different words are associated with different
topics")

# more detailed look at words associated with each topic

# beta values for topic [1]
betaT1<-td_beta %>%
  mutate(topic = paste0("Topic ", topic),
         term = reorder_within(term, beta, topic))
%>%filter(topic=="Topic 1")

# plot word probabilities higher than 0.003 for topic [1]
betaplotT1 <- ggplot(betaT1[betaT1$beta>0.003,], aes(term, beta,
fill = as.factor(topic))) +
  geom_bar(alpha = 0.8, show.legend = FALSE, stat =
"Identity")+coord_flip()+labs(x ="Terms", y = expression(beta),
       title = "Word probabilities for Topic 1")

betaplotT1
```

Understanding the topics using summary visualisations and quoted documents

```
# prints several different types of word profiles, including
highest probability words and FREX words (FREX weights words by
their overall frequency and how exclusive they are to the topic)
labelTopics(model12, n=15, c(11))

# plots this
```

```
plot.STM(model12, type = "labels", topics = c(12), label="frex",
n=10, width=500)

# summary plots
plot(model12, type = "summary", labeltype = c("frex"))
plot(model12, type = "hist", labeltype = c("frex"))
plot(model12, type = "labels", labeltype = c("frex"))
cloud(model12, topic = 1)

# create a data.frame with dropped rows removed to match the
number of documents in stm object
prepped_text_dropped_rows = prepped_text[-c(24, 115, 116, 119,
140, 264, 327, 344, 409, 415, 423, 427, 429, 430, 431, 434, 435,
436, 774, 945, 979, 1204, 1250, 1257, 1429, 1440, 1463, 1484,
1486, 1489, 1513, 1670, 1690, 1706, 1724, 1759, 2311, 2678, 2692,
2698, 2704, 2706, 2708, 2716, 3055, 3371, 3496, 3504, 3541, 3552,
3644, 4362, 4437, 4462, 4601, 4602, 5240, 5244),]

# show a set number of quotations from a specified [topic]
thoughts1 <- findThoughts(model12,
prepped_text_dropped_rows$text, topics=1, n=7)$docs[[1]]

# plot the above
plotQuote(thoughts1, width=150, text.cex=1, maxwidth=500,
main="Topic 1")

# plot the topics in clusters on an interactive web page
stmCorrViz(model12, "corrviz.html",
documents_raw=prepped_text_dropped_rows$text,
documents_matrix=stm_input$documents)
```

Analysis with party as prevalence covariate

```
# bring in party as prevalence covariate and estimate effect on
specified topics [1:12]
prep <- estimateEffect(
  1:12 ~ party + s(year.x),
  model12,
  meta = meta,
  uncertainty = "Global"
  )
```

```
summary(prep, topics = 12)

# plot this
plot(
  prep,
  covariate = "party",
  topics = c(1:12),
  model = model12,
  method = "difference",
  cov.value1 = "Lab",
  cov.value2 = "Con",
  xlab = "More Labour ... More Conservative",
  main = "Estimated Topic Proportion by Major Political Party",
  xlim = c(-0.05, 0.07),
  labeltype = "custom"
  )
```

Analysis with year as second prevalence covariate

```
# create a year sequence to put along the bottom axis
yearseq <- seq(from = as.Date("1970-01-01"), to = as.Date("2021-
05-01"), by = "year")

# plot how topic(s) prevalence changes over time
plot(
  prep,
  covariate = "year.x",
  method = "continuous",
  topics = c(1:12),
  model = z,
  printlegend = FALSE,
  xaxt = "n",
  xlab = "Year"
  )

axis(1,at = as.numeric(yearseq) - min(as.numeric(yearseq)),
labels = yearseq)
seq <- seq(from = as.numeric("1970"), to = as.numeric("2021"))
axis(1, at = seq)
title("Topics relating to strikes and trade unions")
```

```
abline(h=0, col="blue")
```

Analysis with party as topical content covariate

```
# topical content variable allows for the vocabulary used to talk
about a particular topic to vary.
content <- stm(
  docs,
  vocab,
  K = 20,
  prevalence =~ party + s(year.x),
  content =~ party,
  data = meta,
  init.type = "Spectral",
  max.em.its = 75,
  verbose = TRUE
  )

# plot new topics
plot(
  content,
  type = "summary",
  n = 15,
  text.cex = 0.8,
  main = "STM topic shares",
  xlab = "Share estimation"
)

# wordclouds and summaries for each topic
cloud(content, topic = 8)
sageLabels(content, n = c(12))

# plot quotes
content_thoughts3 <- findThoughts(content,
prepped_text_dropped_rows$text, topics=3, n=6)$docs[[1]]
plotQuote(content_thoughts3, width=150, text.cex=1, maxwidth=500,
main="Content Topic 3")

# analyse by party
plot(content, type = "perspectives", n = 40, text.cex=1.2, topics
= 6, covarlevels = c("Con", "Lab"))
```

** Additional Descriptive Analyses

Word frequencies, specifically by party

```
# using unnest function to convert to one word per row tibble +
finding number of words spoken by each party

protest_words <- prepped_text %>%
  unnest_tokens(word, text) %>%
  count(party, word, sort = TRUE)

total_words <- protest_words %>%
  group_by(party) %>%
  summarize(total = sum(n))

protest_words <- left_join(protest_words, total_words)

# removing the same stop words

mystopwords <- tibble(word = c("con", "lab", "one", "house",
"hon", "honourable", "right", "gentleman", "bill", "government",
"bill", "mr", "member", "secretary", "speaker", "home", "prime",
"minister", "friend", "will", "say", "said", "also", "may",
"priti", "patel", "give", "way", "amendment", "clause", "maiden",
"speech", "ms", "constituency", "members", "friend"))

protest_words <- anti_join(protest_words, stop_words,
mystopwords,
                           by = "word")

# plots the words most commonly used by each party

plot_protest <- protest_words %>%
  bind_tf_idf(word, party, n) %>%
  mutate(word = str_remove_all(word, "_")) %>%
  group_by(party) %>%
  slice_max(tf_idf, n = 15) %>%
  ungroup() %>%
  mutate(word = reorder_within(word, tf_idf, party)) %>%
```

```
  mutate(party = factor(party, levels = c("Lab", "Con", "LD"),
exclude = c("PC", "SNP", "Green", "none", NA)))

ggplot(plot_protest, aes(word, tf_idf, fill = party)) +
  geom_col(show.legend = FALSE) +
  labs(x = NULL, y = "tf-idf") +
  facet_wrap(~party, ncol = 2, scales = "free") +
  coord_flip() +
  scale_x_reordered()

# distribution of most common words for each party

ggplot(protest_words, aes(n/total, fill = party)) +
  geom_histogram(show.legend = FALSE) +
  xlim(NA, 0.0009) +
  facet_wrap(~party, ncol = 2, scales = "free_y")

# Zipf's law: rank of each common word by party

freq_by_rank <- protest_words %>%
  group_by(party) %>%
  mutate(rank = row_number(),
         `term frequency` = n/total) %>%
  ungroup()

freq_by_rank %>%
  ggplot(aes(rank, `term frequency`, color = party)) +
  geom_line(size = 1.1, alpha = 0.8, show.legend = FALSE) +
  scale_x_log10() +
  scale_y_log10()

# find the important words for the content of each document by
decreasing the weight for commonly used words and increasing the
weight for words that are not used very much in the whole corpus
of documents

protest_tf_idf <- protest_words %>%
  bind_tf_idf(word, party, n)

protest_tf_idf %>%
  select(-total) %>%
  arrange(desc(tf_idf))
```

```
protest_tf_idf %>%
  group_by(party) %>%
  slice_max(tf_idf, n = 21) %>%
  ungroup() %>%
  filter(party == c("Con", "Lab", "LD")) %>%
  ggplot(aes(tf_idf, fct_reorder(word, tf_idf), fill = party)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~party, ncol = 2, scales = "free") +
  labs(x = "tf-idf", y = NULL)
```

n grams and correlations: finding what words frequently get used together

```
# breaking into 2 word chunks

protest_bigrams <- prepped_text %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2)

# finding most common bigrams

protest_bigrams %>%
  count(bigram, sort = TRUE)

# taking out stop words

bigrams_separated <- protest_bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_filtered <- bigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word)

# new bigram counts:

bigram_counts <- bigrams_filtered %>%
  count(word1, word2, sort = TRUE)

# recombined words after filtering out stop words

bigrams_united <- bigrams_filtered %>%
  unite(bigram, word1, word2, sep = " ")
```

```
# use to see what words frequently get used alongside a
[particular word], by party/year

bigrams_filtered %>%
  filter(word2 == "strike") %>%
  count(party, word1, sort = TRUE)

bigram_tf_idf <- bigrams_united %>%
  count(party, bigram) %>%
  bind_tf_idf(bigram, party, n) %>%
  arrange(desc(tf_idf))

bigram_tf_idf %>%
  group_by(party) %>%
  slice_max(tf_idf, n = 10) %>%
  ungroup()
```

Visualising bigram networks

```
set.seed(2017)

bigram_graph <- bigram_counts %>%
  filter(n > 20) %>%
  graph_from_data_frame()

bigram_graph

# makes visualisation of bigram networks

ggraph(bigram_graph, layout = "fr") +
  geom_edge_link() +
  geom_node_point() +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1)

# more complex visualisation options

set.seed(2020)

a <- grid::arrow(type = "closed", length = unit(.15, "inches"))
```

```
ggraph(bigram_graph, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1) +
  theme_void()
```

Complex bigram/trigram network visualisations using filters

```
# cuts the tibble down into 10 rows at a time, and filters by
year

protest_section_words <- prepped_text %>%
  filter(party == "Lab") %>%
  mutate(section = row_number() %/% 10) %>%
  filter(section > 0) %>%
  unnest_tokens(word, text) %>%
  filter(!word %in% stop_words$word)

# count words co-occuring within sections

word_pairs <- protest_section_words %>%
  pairwise_count(word, section, sort = TRUE)

# search for most common pairing words with a [given word]

word_pairs %>%
  filter(item1 == "traveller")

# Uses the phi coefficient based on how often words co-appear in
a given section

word_cors <- protest_section_words %>%
  group_by(word) %>%
  filter(n() >= 20) %>%
  pairwise_cor(word, section, sort = TRUE)

# gives graphs to show pairings with the highest coefficients
when you input [certain words]

word_cors %>%
```

```
  filter(item1 %in% c("disorder", "demonstration", "protesters"))
%>%
  group_by(item1) %>%
  slice_max(correlation, n = 12) %>%
  ungroup() %>%
  mutate(item2 = reorder(item2, correlation)) %>%
  ggplot(aes(item2, correlation)) +
  geom_bar(stat = "identity") +
  facet_wrap(~ item1, scales = "free") +
  coord_flip()

# puts this into a network visualisation

set.seed(2016)

word_cors %>%
  filter(correlation > .75) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = correlation), show.legend =
FALSE) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), repel = TRUE,
  max.overlaps = getOption("ggrepel.max.overlaps", default =
500)) +
  theme_void()
```