May 2021

Three Essays in Political Economy

Shilpi Mukherjee
Clemson University, shilpionweb@gmail.com

Three Essays in Political Economy

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Economics

by
Shilpi Mukherjee
May 2021

Accepted by:
Dr. Patrick Warren, Committee Chair
Dr. William Dougan
Dr. Chungsang Tom Lam
Dr. Oriana Rachel Aragon

Abstract

This dissertation is comprised of three essays in political economy. In the first chapter, I study the short-run political polarization between Republican and Democrat politicians in the House of Representatives before and after the November 2018 midterm election, using Twitter data. I compute various metrics of ideological polarization at weekly intervals using methods such as hashtag analysis, topic modelling, Bayesian Ideal Point Estimation, mention and retweet network analysis. I empirically check for the patterns in political polarization during the election cycle at the level of discourse. Different measures of polarization signal different patterns in polarization. When polarization is measured by hashtag divergence or topical divergence, it seems to increase as the election approaches. However, when polarization is measured by divergence in word distribution, sentiment-augmented topic divergence, or cited-media ideology divergence, it seems to decrease as the election approaches. This pattern is consistent with a divergence in preferred electoral agenda but convergence in agenda-item-specific positioning.

In the second chapter, I extend the framework of analysis that I developed to the Indian context. I study the short run political polarization between the politicians of the two main national political parties in India contesting in the Lok Sabha, the lower house of the Indian parliament before and after the 2019 general elections, using data from their Twitter feed. I compute various measures of ideological polarization using the methods described in the first chapter, and empirically test the policy convergence hypothesis versus the policy divergence hypothesis discussed in the literature by analysis of these measures of ideological polarization. This chapter reiterates the findings of the previous chapter, which shows that the different measures of polarization signal different patterns in polarization. We find increase in polarization as measured through topical divergence and fall in polarization as measured through sentiment augmented topic analysis suggesting divergence in agenda setting behavior and convergence in agenda-item-specific positioning. This is similar to the pattern in the U.S data. However, in contrast to the US data, polarization as measured through hashtag divergence decreases whereas polarization as measured by cited media ideology increases as we approach the election in India.

In the third chapter chapter, my coauthor Sagnik Das from City University of New York and myself study the effect of political business cycles on government expenditure in India as measured using data from the world's most extensive public works programme (NREGA), new road constructed data as well as night light intensity data which is used as a proxy for development. Using panel data at the district level spanning from 2011 to 2020 for NREGA employment and expenditure, 2000 to 2014 for new road constructed under the PMGSY program and mean total calibrated night light intensity from 1994 to 2014, we can show the existence of political business cycles wherein politicians stimulate the economy before the election either to lure myopic voters or to signal their capability to forward-looking voters. We find the causal impact of political business cycle on expenditure undertaken under NREGA and on employment provided under NREGA at the intensive margin. We also find evidence of political business cycles impacting the length of new road constructed under PMGSY and money disbursed by the Government for new projects to be undertaken under PMGSY. For night light intensity, we do find some evidence of the causal impact of political business cycles. We also use high-frequency monthly night light intensity data spanning from 1993-2013 to investigate the political business cycle's effect in the shorter run. We do see a statistically significant spike in night light intensity one month before the election. However, we are unable to find any conclusive trend with the approach of the election.

Dedication

To my parents for their unconditional support and encouragement and my brother for his constant companionship.

Acknowledgments

I would first like to thank my advisor Dr. Patrick Warren who has always been a source of inspiration, encouragement and unfaltering guidance. He has always let me pursue my ideas to the fullest while always correcting my mistakes and providing fresh insights. I thank him for aiding in my development as an independent researcher. I am also extremely thankful to the members of the Wednesday Warren Workshop, who provided their feedback on my work, and helped me improve my papers. I also learnt a lot from their research which I believe has helped me become a better economist.

I am also extremely thankful to my committee members Prof. William Dougan, Prof. Tom Lam and Prof. Oriana Aragon who have been instrumental in providing me guidance on various stages of my work. They always kept their doors open for me and provided me with feedback and suggestions that has helped me develop my project to its current version.

I am particularly indebted to the members of the Public Economics Workshop and IO Workshop who have helped me streamline my research question, improve versions of my paper and in general contribute to the development and progress of my research. I thank Dr. Frederick Hanssen, Dr. Charles Thomas, Dr. Matthew Lewis, Dr. Christy Zhou, Dr. Michael Makowsky, Dr. Robert Fleck, Dr. Howard Bodenhorn, Dr. Molly Espey, Dr. Arnold Harberger, my committee members, and my peers at the workshops.

Thanks to the faculty and administrative staff at Clemson University . Thanks to Dr. Scott Baier, Dr. Curtis Simon, Kimberly Albertson, Jill Watkins and Lisa Doyle for helping me throughout my PhD career. I would also like to thank the Graduate Student Government at Clemson University who had made my travel to conferences easier through their generous funding opportunities and also CCIT for providing generous computing resources without which this research would not have been successful.

I also want to express my gratitude for my professors at Calcutta University especially Dr. Sukanta Bhattacharya and Dr. Anirban Mukherjee who were my advisors for my MPhil thesis and Dr. Debabrata Lahiri from my undergraduate institution, Presidency College for supporting me in my desire to pursue a graduate education and providing the bedrock for my economic fundamentals. I am also thankful to my school teachers Ms.Sumita Roy Chowdhury and Mr. Arijit Ghosh and Mr. Abhishek Bagchi for believing in me and motivating me.

I would want to acknowledge my peers in school Shannon, Jeremy, Adam, Bob, Majid and especially Aspen who made me feel so welcome in the country and were great to share a classroom with.

Last, I would want to thank my friends and family without which this thesis would not have been possible. My friends Maneesh, Sagnik, Priyanka and Pardhvi were more than family in a foreign land and stood by me through thick and thin. I also want to thank Ankur, who has stood by all my quirks and mood swings with a happy face and encouraged and motivated through every trough I have faced.

In the end, I am forever grateful to my parents for just being themselves, for supporting me through all my decisions and for raising me independently and my brother for being one of my closest confidante, and for making things easier for me whenever possible.

Thank you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Does Congressional Polarization Decline as Election Approaches: Evidence from Twitter Data in USA

1.1 Introduction

This paper presents novel high-frequency estimates of partisanship constructed using Twitter data as applied to the US representatives. Political partisanship affects not only preferences and voting behaviour but also other political and economic outcomes. Survey results from Gerber and Huber (2009) show that beliefs about current and expected future economic performance are more positive when the respondent's partisanship matches that of the current President. McConnell et al. (2018) show that partisanship can also affect non - political markets such as the labour market and the goods market. Using a sizeable experimental design, they show that workers are willing to take a lower pay from co-partisans, suggesting a compensating differential. In the market for goods, the authors find that buyers are nearly twice as likely to engage in a transaction when they and the seller share the same partisanship.

Although political partisanship has been shown to affect key outcomes, the theory of political partisanship is mixed. Several papers have modelled a theory of short-run political partisanship where parties are only motivated to win elections (Downs (1957), Comanor (1976), Ledyard (1984) and others). In these models, where parties are assumed to be only motivated to hold office, their proposed equilibrium policy platforms converge before the elections. However, the policy convergence hypothesis has been subjected to a lot of subsequent scrutiny, which has been critical of the hypothesis and predicted policy divergence (Alesina (1988), Coughlin (1992), Besley and Case (2003), Glaeser et al. (2005) and others).

On the other hand, research on long-run political polarization in the U.S. has shown that polarization between parties has been on the rise since World War II. Although there have been arguments about whether ideological scores are just moving to the pre-WWII era or if there is real ideological polarization, there is some consensus that polarization has been increasing for the last few decades. Figure 1.1 shows the distance between the mean ideological score for the Democrat and Republican politicians based on the first dimension of the DW-Nominate scores.[1] The first dimension computes the ideology of politicians along the liberal-conservative scale. The figure shows that polarization in the House and the Senate has been increasing since the 1980s, and is at an all-time high in the past 125 years.

[1] The trend looks very similar when we use the Poole and Rosenthal estimate for the first dimension

There also has been some research on how partisanship has evolved in the medium run by leveraging text data. Gentzkow et al. (2019) show that partisanship, as measured by Congressional speeches, is much more significant in the recent years than the past, and has increased substantially since the 1990s, following the Contract with America, Gingrich et al. (1994), remaining relatively flat before that. Several articles have also shown that partisan differences have seeped into media language and of late the two parties are using strategically different languages through consultants, focus groups and polls. Lakoff (2003) suggests that this represents a substantial change in how partisanship has evolved in recent years, especially with respect to linguistics.

Therefore, substantive empirical research shows rising rates of political polarization between Democrats and Republicans in the long run and the medium run. However, due to data paucity, there has not been substantial empirical research on short-run political polarization, even though theoretical research abounds. A potential reason for the dearth of empirical analysis of short-run political polarization is that no good secondary source of data is available for measuring political ideology, except for the DW-Nominate scores. These scores are constructed using roll-call data. However, roll-call data is only available for each session of Congress, and therefore cannot be used to understand ideological movements during the election cycle. Survey data can help us look at polarization in a time series fashion at a high granularity level. However, collecting survey data at such short intervals is very costly, especially for elites such as Representatives or Senators. To circumvent this problem, I look at short-run polarization between Democrat and Republican politicians using their Twitter feeds. The Twitter data helps to zoom in on the politicians' behaviour in the short run, at a high granularity level.

To quantify the degree of polarization between Republican and Democrat politicians, I compute several metrics to measure expressive ideological polarization using Twitter data. I use the term ‘expressive ideology’ because I measure their ideological estimates using Twitter’s rhetorical data and not their actual behaviour. To do this, I collect tweets from incumbent politicians in the 116^{th} House of Representatives, one year before and after the midterm elections conducted on 6^{th} November 2018 at a weekly interval. I exploit both the tweets’ linguistic aspects and the network structure of the tweets to compute several metrics of polarization. For understanding the linguistic aspect, I perform Hashtag Analysis, topic modelling using Latent Dirichlet Allocation(LDA) and Bayesian Ideal Point Estimation to compute various estimates of polarization. I also study the retweet and mention networks of politicians and construct measures of polarization based on these networks.

A brief definition of each of these metrics is given here. Hashtag Similarity measures the number of common hashtags used by Democrats and Republicans conditional on the top hashtags used by them. Inverse of Sentiment Augmented Hashtag Distance ($Inverse\_Score\_std$) is the inverse of the euclidean distance between vectors of fractions of tweets corresponding to a particular hashtag with a negative sentiment by Republicans and by Democrats. The lower the $Inverse\_Score\_std$, the lesser is the polarization. Hellinger Distance and Kullback-Leibrar Divergence measure the distance between topic distributions (distributions of topics used by Democrats and Republicans in their tweets) whereas the Jaccard Distance measures the distance between word distributions (distributions of words used by Democrats and Republicans in their tweets). Next, each tweet is assigned to the most dominant topic contained in the tweet. Dominant Topic Distance ($Score$) calculates the Euclidean distance between two vectors of fractions devoted to each topic by Democrats and Republicans. Sentiment Augmented Dominant Topic Distance (Sum_Frac_Dis) measures the euclidean distance between the vectors of fractions of tweets that Republicans and Democrats devote to positive, negative and neutral sentiment for each topic. The greater the distance, the higher is the polarization for these distance-based measures. I also perform Bayesian Ideal point estimation based on the politicians' URL (Uniform Resource Locator) sharing behaviour to estimate the politicians' ideologies. For the mention and retweet network analysis, I calculate the share of how many times a Democrat mentions/retweets a Republican negatively and vice versa. One can find a detailed discussion on the construction of these metrics in Section 1.5.

Polarization, as measured by these metrics, varies considerably in the election cycle. To get a sense of what happens close to the election, I zoom into eight weeks before and after the election to look at what happens as we move into and away from the election. The choice of 8 weeks is made because there are no primaries in this period, and I hope to capture the upcoming midterm election effect. I report the broad trends in these metrics here. Since there are only eight weeks of data, these metrics' slope estimates are not very precise, and most of them are not statistically significant. One way to interpret these results strictly from the perspective of statistical significance would be to say that there is no evidence of a decrease or increase in polarization as we approach or move away from the election. This suggests that the theoretical literature of policy convergence versus divergence does not play out in politicians' Twitter feeds. However, since we only have one year of data, and if we are willing to take a Bayesian approach, looking at the coefficients does suggest some patterns.

Polarization as measured by hashtag similarity and Inverse of Sentiment Augmented Hashtag Distance ($Inverse\_Score\_std$) increases as we approach the election, falls once the election is over, and increases over the next eight weeks after the election. The distance between topic distributions also increases as we approach the election, falls once the election is over and keeps falling over the next eight weeks. Dominant Topic Distance (*Score*) increases as we approach the election, falls after the election, and keeps falling over the next eight weeks. On the other hand, Jaccard distance which is the euclidean distance between word distributions falls as we approach the election, falls after the election and increases as we move away from the election. Sentiment Augmented Dominant Topic Distance (Sum_Frac_Dis) falls as we approach the election, increases once the election is over and falls again. The ideological difference, as measured by the difference between Bayesian Ideal Point Estimates falls as we approach the election, falls once the election is over and then keeps increasing as we move away from the election.

I find that there is an implicit and explicit element to polarization. As measured by hashtag analysis, topic divergence and retweet and mention networks, polarization increases as we approach the election suggesting that politicians get more polarized in their agenda-setting behaviour. However, negative retweets of Democrats by Republicans decrease in the last eight weeks in the approach to the election. On the other hand, more implicit measures of polarization, as measured by divergence in words used, sentiment augmented content analysis, and URL sharing behaviour decreases as we approach the election. It makes sense to think that politicians are probably trying to appeal to their electoral bases through their agenda-setting behaviour regarding the hashtags that they use and the topics they talk about. However, there is a decrease in polarization in terms of the conversations within a topic. This suggests that politicians are trying to appeal to the median voter through the content within a topic. Therefore, we find that politicians can and do use different aspects of a tweet to talk to different sub populations and while they might seem to diverge in the agendas that they talk about they might be converging within a particular agenda.

My contribution in this paper, therefore runs along three dimensions. First, I help test the policy convergence versus policy divergence hypothesis, using a very rich data set that allows me to compute ideological scores from various aspects of Twitter data. This is the only paper to compute high-frequency ideological estimates of polarization in a time-series fashion. Second, I add to the methodology of measuring political polarization in the literature using the distance between two topic distributions and sentiment augmented content analysis. Therefore, these methods can be replicated in a less developed country fairly easily, which might not have official data like roll call votes found in the U.S.[2] Third, I also show that various dimensions of a tweet, such as hashtags, content, networks which can all be used to convey information need to be studied separately as they sometimes can provide competing signals.

The rest of the paper is organized as follows. Section 1.2 talk the relevant theoretical background briefly. Section 1.3 discusses other relevant literature. Section 1.4 discusses the data. Section 1.5 discusses the methodology of computation of the various measures. Section 1.6 talks about the empirical methodology and the results and Section 1.7 concludes.

1.2 Theoretical Background

There have been broadly two approaches to modelling short-run political partisanship. The first way short-run political partisanship has been modelled assumes

[2] Although the algorithms are scalable, they are costly in terms of computational and manual resources.

that political parties are only interested in holding office or winning the election. The most seminal result from this assumption, as shown by Downs (1957) assuming rational voters and using Hotelling (1990) model of spatial competition is that parties converge on the policy preferred by the median voter. Comanor (1976) has shown that the median voter theorem holds under reasonable degrees of skewness of political preferences. Other works such as Ledyard (1984), Coughlin and Nitzan (1981), Hinich (1976) also predict policy convergence of other types, though not always to the preference of the median voter, in the case of office motivated political parties.

Other papers consider that politicians or parties which are sometimes considered synonymous are ideologically motivated. Wittman (1983), Calvert (1985) consider ideologically motivated politicians, but predict that equilibrium policies chosen by the parties are very close to each other because they assume that parties have a binding commitment to their policy platform and assume certainty about voters' preferences.

There have been several ways that scholars have challenged the policy convergence hypothesis. Alesina (1988) in his seminal paper shows that in the absence of a binding commitment device, which is mostly absent in elections, in a one-shot electoral game, political parties have no incentive to stick to their announced policy level, after the election is over. Rational voters can correctly anticipate the politicians' behaviour, and therefore expect to have policy divergence, which then becomes the equilibrium strategy for the political parties. If the parties have a reasonably high discount factor, then in an infinitely repeated game, they might sustain a convergent policy position. However, if one party believes that the other party's leadership may change, leading to a low discount factor, the co-operative equilibrium breaks. The parties revert to the outcome of the one-shot electoral game predicting policy divergence. Some papers assume uncertainty about voters' preferences by the candidates Coughlin (1992) and Ordeshook (1986) making them unsure of how voters will choose and thereby predicting platform divergence. Other models of this flavour assume uncertainty about the degree to which candidates will implement campaign promises, non-policy considerations like candidates' personal qualities, imperfect mapping of candidates' positions due to incomplete information, weighing of decisions by candidate "competence," ambivalence towards candidates' positions, or other unpredictable factors such as voter mistakes Erikson and Romero (1990), Bartels (1986), Alvarez and Brehm (1995).

Apart from the commitment problem and problem of uncertainty of preferences, Osborne and Slivinski (1996), Besley and Coate (1997) develop citizen candidate models, where citizens who are ideologically motivated contest in an election and implement their preferred policy on winning. These models also predict policy divergence. Several papers also use the probabilistic voting model where elites are uncertain about voter's preferences. Glaeser et al. (2005) in their paper talk about another reason for policy divergence. They argue that if promoting extreme party positions, helps in the sorting of voters such that it increases donations and voter turnout, thereby increasing the probability of winning, parties will prefer to diverge their policies.

Therefore, my aim in this paper is to look at the polarization between politicians in the context of social media over a shorter time period and at a high frequency. According to the median voter theorem, politicians should converge in expressed ideology/proposed policy as an election approaches. Since traditional data on politicians ideology is measured by DW-Nominate Score, which is based on Congressional voting records, it is not possible to see the movement of ideological scores over short intervals of time, due to data constraints. To solve this problem, and to understand elite polarization in the context of social media, I use Twitter data from the incumbent Congressional members of the 116th House of Congress and compute metrics of their ideology using Twitter data from one year before and after the election and see if their expressed ideologies converge as the election on November 6, 2018 approaches.

1.3 Other Relevant Literature

Juxtaposed with the theoretical and empirical modelling of short-run political partisanship, another literature has documented the growing long-run polarization between Republicans and Democrats. Aldrich (1995), McCarty et al. (1997), Jacobson (2000), Hetherington (2001), Collie and Mason (2000) have all shown the growing political polarization between Democrats and Republicans in the U.S in the government. Layman et al. (2010) show that Republicans and Democrats have become polarized both in the government and in the electorate through conflict extension along several dimensions.

In the backdrop of increasing long-run political polarization, I compute high-frequency estimates of political polarization in the short-run. Twitter data is an excellent source of data to identify this kind of variation in ideology for several reasons. Twitter data helps us gain insight into the political discourse at very short intervals, almost week by week. A possible flip side is that Twitter data, unlike roll call data, does not give us any access to behavioural insights regarding actual votes in favour or against a particular policy. However, the fact that politicians have increasingly taken to Twitter to state their policy positions, lends legitimacy to analyze their Twitter feeds. For example, the former President Donald Trump sent out 518 tweets (11 deleted) in the first 100 days of his Presidency according to politico.com, meaning that he alone sent out five tweets on average per day. This is not a trait specific to the former President. Most politicians have tried to make use of the "Obama Model" to reach the general public, Towner and Dulio (2012). The way that social media has been used for political campaigning has been documented in a number of research studies (Adams and McCorkindale (2013), Conway et al. (2013), Golbeck et al. (2010), Graham et al. (2013), Grant et al. (2010), Johnson and Perlmutter (2010), Xiong et al. (2019)). These studies testify that ever since Twitter came into effect in 2006, it has been increasingly adopted by politicians worldwide to influence their campaign strategy. Social media and traditional media are also found to have a symbiotic relationship in terms of agenda-setting during election campaigns as found in a paper by Conway et al. (2015). They investigated the relationship between the Twitter feeds of political candidates and parties and the news media output.

Although social media has been considered an essential element of political campaigning, social networking sites (SNSs) such as Twitter are considered the favoured forms of social media for campaign purposes. SNSs are unique because they allow connections to be displayed openly. However, unlike some SNSs which have privacy controls, Twitter users have mainly public profiles, which do not require bidirectional confirmation of networks (Boyd and Ellison (2007), Vergeer (2015)). This allows it to be used as a broadcast medium, an extensively used attribute in political campaigning. This makes Twitter a very natural choice for the question that I attempt to answer.

1.4 Data

1.4.1 Twitter Data

The United States of America had a midterm election on 6th November 2018 where 435 seats from the US House of Representatives were contested. I collected the official Twitter handle of all the incumbents from the US House of Representatives who also won in the 2018 elections. This was done by searching for verified handles for each of the incumbent members. In the presence of more than one verified handle, I considered the one which had the link to the Representative's official page in the official website of the US House of Representatives.[3] I collected tweets at weekly intervals for each politician tweeted by them from 7th November, 2017 to 5th November, 2019, a total of 104 weeks or two years of data using an application called Social Studio.[4] For some of the weeks where data was not available through Social Studio, I used the Twitter API.

In my data set, I have access to the name of Congressperson who tweeted, the tweet's content, the publish date and time of the tweet and whether the tweet was a normal tweet, retweet or quote tweet.

There are 177 Democrats, and 165 Republicans in my data set[5]. Only one Democrat incumbent winner and two Republican incumbent winners did not have an official handle. The number of politicians in my data set account for 78.62 % of the total

[3] This was true for famous politicians, whose campaign accounts or personal accounts were also verified.

[4] I am thankful to the Social Media Listening Center, Clemson University for providing me access to the Social Studio app.

[5] The twitter handles were collected in October 2019, and the Congresspeople who had verified twitter handles then, are included in the data-set.

number of politicians in the U.S House of Representatives.[6] Figure 1.2 shows the number of active incumbent politicians over time, with the number being calculated every week. A politician is defined as an active politician even if s/he makes atleast one tweet in the span of the week, under consideration, s/he is counted into the number of active politicians. So, if politician X tweets one message in the 22^{nd} week, but not in the 23^{rd} week he/she would be counted as an active politician in the 22^{nd} week but not in the 23^{rd} week. We have an average number of 146 active Republicans and 168 active Democrats in our data set every week. The higher number of average active Democrats compared to Republicans is attributed to two reasons. First, there are higher number of Democrats in my data set, but second whereas approximately 94.91 % of the total Democrats are active authors, 88.48 % of Republicans are active authors.

Figure 1.3 shows the total tweets by Democrats and Republicans over time, computed every week. This figure shows that based on absolute numbers, Democrats tweet more than Republicans. Republicans post around 1852 tweets every week, whereas Democrats post 3350 tweets every week. To see if this variation comes only from the higher number of Democrats in my sample or if Democrats demonstrate a higher tweeting propensity than Republicans, we refer to Figure 1.4. Figure 1.4 shows the average number of tweets by each active incumbent member of the House over time, computed every week. This figure shows that even after controlling for the number of active authors, Democrats tweet more on average than Republicans. Whereas a Republican posts 13 tweets on average per week, a Democrat posts 20 tweets on average per week.

[6] I only consider incumbents who contested in 2018 and won, so that I can get their tweets from the official handles after the election too. However, there were only three incumbents who contested in the 2018 election and did not win, and hence I argue that since the number is so small, it should not introduce a huge selection bias in my data

Figures 1.2, 1.3, 1.4 also show some broad trends. For example, for all the graphs, we see a dip in the numbers during the last week of December and the beginning of New Years, when politicians are probably spending time with their families. Another unique thing about the figures is that Democrats and Republicans' tweets' trend appears to follow the same pattern in the crests and troughs. Whether this is due to some underlying causal factor or a feedback mechanism between Democrats and Republicans is unclear.

1.4.2 Competitiveness of Race Data

Along with the Twitter data, I also collect the competitiveness of race information for all the congressional districts of the United States for the 104 weeks that are there in my data-set. I collect this data from the Cook Political Report by scraping the website for the data. The data provides information on whether a district is Solid Republican, Solid Democrat, Likely Republican, Likely Democrat, Lean Republican, Lean Democrat, Republican Toss-Up or Democrat Toss-Up. As the name suggests, solid refers to the safest districts, followed by Likely and Lean, whereas Toss-up refers to the most competitive districts. There are 64 weeks of unique data. I match up the competitiveness of race data with the 104 weeks of data in my original sample, by assigning the value of race competitiveness of a particular district in a particular week to the closest race competitiveness data available at that time.[7] This allows for time-sensitive data on race-competitiveness. Auter and Fine (2016) use this measure of the competitiveness of an election in their paper on negative campaigning on Facebook. They find that underdog candidates in less-competitive races indulge in

[7] There are 64 weeks of race competitiveness data in my data-set. There is no competitiveness data between 5th November 2018 and 12th April 2019. I assign the race competitiveness data of either 5th November 2018 or 12th April 2019 to the weeks for which I do not have any race competitiveness data depending upon which date is closer to the particular week in question.

negative campaigning in issue attacks, whereas candidates in competitive races are more into a personal attack. There is, therefore, a reason to believe that the competitiveness of race in a district will also influence the ideology of Democrats and Republicans in that district.

1.5 Computation of Ideological Estimates

To understand whether politicians behave differently at least in the domain of rhetoric in the advent of an election, I analyze the tweet's content from a linguistic perspective, along with the network structure of the tweeting behaviour of the politicians. With the increase in computational power and the explosion in unstructured data, text-data analysis is continually being used to answer various questions and is being considered an increasingly important data source, Gentzkow et al. (2019). Gentzkow and Shapiro (2010) use text data to develop an index of media slant to assess the similarity of the language used by a news outlet to that used by a Republican or Democrat. Social scientists have also analyzed text data for understanding polarization specifically. Gentzkow et al. (2019) study the partisanship trend in Congress by analyzing speech from 1873 to 2009. Ash et al. (2017) similarly look at U.S Circuit court judges' polarization using text data of the court opinions from the 1890s to 2010s. Bara et al. (2007) analyses parliamentary debates in the U.K to identify the dominant themes in debate and also the difference in discourse between leaders favouring different policy positions.

Some other studies have specifically used Twitter to understand polarization. Demszky et al. (2019) use the natural language processing framework to understand political polarization in Twitter, in the context of 21 mass shootings in the USA. Monti et al. (2013) model political disaffection using Italian Twitter data by employing sentiment analysis.

Although the use of text data is increasing over time, the field itself is in its early stages and is still evolving. There are several different methods available that scholars have used in the past. In this paper to analyze the linguistic aspect of the tweets, I perform Hashtag analysis, topic modelling, sentiment analysis, and Bayesian Ideal Point estimation to create metrics of ideological polarization. In assessing the network structure of the politicians' tweeting behaviour, I study the retweet network and the mention network complemented with sentiment analysis of the tweets.

1.5.1 Hashtag Analysis

To start with the computation of metrics of polarization, I look at the similarity in hashtags used. Hashtag similarity is defined as the number of common hashtags used by Democrats and Republicans conditional on the top hashtags used by them. To compute the hashtag similarity, I proceed in the following way: First, I extract the top 40 hashtags used by Republicans in a week. Let us denote this set of hashtags by R_{40}. Second I extract the top 40 hashtags used by Democrats in a week.[8] Let us denote this set of hashtags by D_{40}. I then compute the number of similar hashtags between the sets R_{40} and D_{40}. Let us denote this by $Hashtag_{40}$. In other words,

$$Hashtag_{40} = n(R_{40} \cap D_{40}), \tag{1.1}$$

where $n(.)$ denotes the cardinal number. I also compute $Hashtag_{10}$, $Hashtag_{20}$, $Hashtag_{50}$, $Hashtag_{100}$ for robustness checks. These values show how the usage of

[8] I convert all the hashtags to lower case because sometimes the same hashtags can be written in different cases.

common hashtags used by Republicans and Democrats vary conditional on the top hashtags used by each group. Table 1.1 for example shows the top 20 hashtags one week before and after the election. The italicized and underlined words show the common hashtags used both by Republicans and Democrats. In the week before the election there is only one common hashtag whereas in the week after the election, there are four common hashtags.

Hashtags in Twitter are used as an organic and community-driven method to add context to the data, Wang et al. (2011). They can, therefore, be thought of as broad topics that the Twitter users are talking about. Some studies have also shown that hashtags are sometimes used as framing devices, Moscato (2016) in guiding the political conversation. Bruns and Burgess (2015) talks about how hashtags have evolved from ad hoc devices in Twitter to tools that can be used to organize movements and guide the discussion of topics in the platform. The role of hashtags in guiding social and political movements have been studied in many situations such as Canadian elections, Arab Springs movement, student protest movement against high fees in Africa, and the recent feminist movement which is best known by the hashtag it used, #MeToo (Langa et al. (2017), Small (2011), Moscato (2016), Huang (2011), Bruns et al. (2014)). Therefore, I start by looking at the similarity of hashtags used by Republican and Democrat politicians over time as they give us the first piece of evidence of the way conversation changes between these two groups as the election approaches and if it changes once the election is over.

Figure 1.5 shows the trends of these metrics over the election cycle. According to this figure, the overlap between hashtags keeps decreasing as the election approaches, and increases after the election. This suggests that Democrats and Republicans talk about different agendas in the approach to the election. The same pattern is valid for all four trends. I also do the hashtag analysis for tweets which have a negative sentiment and a positive sentiment and get similar results. This means that irrespective of the tweet's sentiment, the number of common hashtags used by Republicans and Democrats fall as they approach the election. The figures are presented in the Appendix.

1.5.1.1 Sentiment augmented Euclidean Distance between hashtags

Hashtags are generally very context-specific and are used to convey only a particular sentiment, as shown by the hashtags in Table 1.1. While Democrats use #getcovered, #protectourcare and #goptaxscam, Republicans use #taxreform, #taxcutsandjobsact and #maga. However, there can be times when a particular common hashtag is used positively by representatives from one party but negatively by representatives from the other party. In that case, an increase in the number of common hashtags might give us a false sense of decreasing rhetorical polarization between the two parties. To tackle this problem, I perform a sentiment analysis of the tweets containing hashtags. I use the Vader Sentiment analysis module in Python, which is a valence based sentiment analysis module developed especially for micro-blogging sites such as Twitter, Hutto and Gilbert (2014).[9] The package computes the positive, negative and neutral polarity for each tweet. It also gives a compound score. If the compound score is less than -0.5, the tweet is considered negative, if it is greater than +0.5, the tweet is considered positive, and if the scores lies between -0.5 and + 0.5, the tweet is considered to be neutral.

[9] They use a gold standard of lexical features as well as the polarity and intensity of words to compute the sentiment score. They also show that their approach is better than eleven of the common and most widely used Sentiment Analysis methods and outperforms human accuracy.

To combine the hashtag analysis with the sentiment analysis, I compute the distance between the fraction of negative tweets for the Republicans and Democrats out of all the tweets that use a similar hashtag. For example, a common hashtag in the top 40 hashtags used by Democrats and Republicans for the week of 1st August to 7th August 2018 is #smallbusinessweek. I count the fraction of tweets made using a negative sentiment using the hashtag #smallbusinessweek by Democrats and Republicans. I repeat this for all the common hashtags in $Hashtag_{40}$ and calculate the Euclidean distance between those two vectors.

In other words, assume that there are s common topics in the top 40 hashtags used by Republicans and Democrats. Therefore, the length of $Hashtag_{40}$ which we have already defined is s. I now construct two vectors D_{40_s} and R_{40_s}. Let the first element of D_{40_s} be denoted by $D_{40_s}(1)$. Then,

$$D_{40_s}(1) = \left[\frac{n\begin{pmatrix} Tweets\ by\ Democrats\ which\ contain\ the\ first\ hashtag \\ \hfill in\ top\ 40\ hashtags\ and\ have\ a\ negative\ sentiment \end{pmatrix}}{n\begin{pmatrix} Tweets\ by\ Democrats\ which\ contain\ the\ first\ hashtag \\ \hfill in\ top\ 40\ hashtags \end{pmatrix}} \right], \tag{1.2}$$

where n denotes the cardinal number. I similarly compute all the elements for D_{40_s}, R_{40_s} and find the euclidean distance between these two vectors. This is denoted by $Score_{40}$, where $Score_{40}$ is defined as follows:

$$Score_{40} = d(D_{40_s}, R_{40_s}) = \sqrt{\begin{matrix} (D_{40_s}(1) - R_{40_s}(1))^2 + (D_{40_s}(2) - R_{40_s}(2))^2 \\ \hfill + ... + (D_{40_s}(s) - R_{40_s}(s))^2 \end{matrix}}. \tag{1.3}$$

After this, I standardize the scores by the number of common hashtags by dividing

$Score$ by the square root of the number of common hashtags. For example, $Score_{40_{std}}$ is computed as follows

$$Score_{40_{std}} = \frac{Score}{\sqrt{s}}. \tag{1.4}$$

I similarly also compute $Score_{10_{std}}$, $Score_{20_{std}}$, $Score_{50_{std}}$ and $Score_{100_{std}}$. I focus only on negative sentiments because hashtags, being context-specific, have either a positive or a negative undertone. Hence, it does not make much sense to distinguish between tweets with a positive sentiment and a neutral sentiment in case of hashtags. Therefore, treating positive and neutral tweets as a single non-negative category essentially means that we only need to calculate the negative category's Euclidean Distance. A high Euclidean Distance in the negative category automatically implies that distance in the non-negative category is also high and vice-versa.

Another point to note is that in doing the actual analysis, I use the inverse of $Score_{10}$ and $Score_{10_{std}}$ which I refer to as Inv_Score_10 and $Inv\_Score\_10\_std$. This is done because if there are no common hashtags for any of the groups, then the distance would be calculated as 0, but that does not make sense. A distance close to 0 implies no polarization, whereas 0 common hashtags do not imply the same. To resolve this ambiguity, we take the inverse of the score, such that a high score means less polarization and low score means high polarization. When there are no common hashtags, the metric is set to a value of 0, as no common hashtags imply the most significant degree of polarization.

Fig 1.6 shows the trend of the inverse of the standardized scores. The non-standardized scores look the same and are included in the Appendix. This graph shows that as we approach the election, the inverse of the euclidean distance falls, or in other words, the euclidean distance increases. This means that conditional on using the same hashtags in the top hashtags that Democrats and Republicans use; they use it with different sentiments as they approach the election implying that polarization in using hashtags increases as the election gets closer. The trend after the election is different for top 10 and 20 hashtags versus the others. This might be because some hashtags in the top 10 or 20 hashtags have very different characteristics compared to the hashtags in top 40, 50 or 100. Nevertheless, it is quite clear that the euclidean distance has a clear pattern before the election, and it increases as we approach the election.

1.5.2 Topic Modelling

After looking at the hashtags, which are broad level agenda setting items, I look at the tweets' content directly. This allows us to understand the data even better. To do this, I use the method of topic modelling. Topic modeling is an unsupervised machine learning algorithm that scans through a set of documents, detects word and phrase patterns within them, and automatically clusters words and phrases within those documents. These documents can be news articles, congressional speeches, parliamentary debates or in my case tweets. Topic Modelling helps us go one step further in looking at the divergence between Republicans and Democrats by looking at the tweets' content. A number of very prominent and influential studies have been conducted in the field of information retrieval and automatic detection of topics in political speeches (Steyvers et al. (2004), Mamou et al. (2007), Quinn et al. (2010)). Boyd-Graber et al. (2017) shows the recent topic modelling applications for information retrieval, linguistic understanding, statistical inference and other tasks. Topic modelling has also been used in the domain of social media data. Lucas et al. (2015) analyses how to perform topic modelling for tweets in different languages.

For applying topic modelling to my case, I use the tweets' content sent out by the politicians.

I apply the Latent Dirichlet Allocation (LDA) model to my corpora of tweets to perform topic modelling.[10] One of the parameters that need to be provided to the LDA model is the number of topics in the corpora. There is no perfect objective measure to estimate the optimal number of topics for a given corpus, in the literature yet. One of the ways to estimate the right number of topics, is to look at the coherence score, for the different number of topics, and select the number of topics, when the coherence score stops increasing.[11]

The LDA algorithm, first developed by Blei et al. (2003), has revolutionized information retrieval. LDA is an unsupervised, probabilistic machine learning algorithm that automatically groups words based on which words occur together more frequently in a corpus of data. Barberá et al. (2018) uses LDA model on tweets by the 113th Congress members, select media outlets, and other groups of people, such as general public, attentive, close party supporters, media and show using a Vector Auto Regression model that politicians are most attentive to issues of close party supporters

[10] To apply the LDA model the data needs to be pre-processed in order to be ready for the application of Latent Dirichlet Allocation (LDA). In keeping with the norms of Natural Language Processing (NLP), and the specificity of Twitter data, I remove the special characters such as '@', '#' specific to Twitter, and punctuation like the period, comma, semicolon and others. I also remove all the stopwords, words such as 'the', 'in', 'from', etcetera, which are very common in the English language, but devoid of any meaning.I use the NLTK corpus of stopwords for the English for this step of pre-processing. I then lemmatize the data, which means that all words in our data-set, (referred to as tokens in the NLP nomenclature), are converted to their base form. For example, am/is/are are all converted to be. I also stemmed the data, which is another form of converting the words to their base form. However, in the context of Twitter data, lemmatization seems to be better at tokenization than stemming. I also create bigrams and trigrams to capture words that might always be associated together. For example, the term 'White House' is an excellent example of a bigram, that could be present in our data-set. We would lose the significance of the term "White House", if we used only the unigram model, which would treat 'white' and 'house' as two separate words.

[11] Since, the LDA is a probabilistic model each run of the model, generates new values of the coherence score. I set a random seed equal to zero so that the model can be replicated.

for setting the agenda. Nardi Jr (2012) uses the LDA model to analyze Supreme Court Decisions' text in the Philippines Supreme Court. Jacobi et al. (2016) uses the LDA model to study large volumes of journalistic text from The New York Times from 1945 to the present. Sokolova et al. (2016) identified election related events using LDA. Ryoo and Bendle (2017) use the LDA model to study the social media strategies of the two campaigns in the 2016 U.S election. The model can be used to infer what percentage of each topic is present in a particular tweet. This helps us understand which topic a particular tweet is about. Figures 1.7 and 1.8 show the first topic in a LDA model fitted over Republican and Democrat politicians' tweets separately with ten topics, in the month of November to December 2017. The figure shows that the first topic mostly deals with taxation and economy, whereas we see words such as *taxreform*, *economy*, and *american* in the tweets by Republicans, and words such as *goptaxscam* and *middleclass* by the Democrats.[12]

To implement the LDA model for my data, I first run the model on each week of tweets for both Republicans and Democrats combined.[13] I run the LDA model for single-digit topic numbers because using more than those number means that my topics are going to have a sparse number of tokens/words, and more than nine topics seem too many for one week. I calculate each of these models' coherence score and choose the optimal number of topics based on the coherence score. The model is then re-run with the optimal number of topics for the Republican tweets and Democrat tweets separately. This provides us with two topic distributions, one for the Democrats and one for the Republicans. After running the optimal LDA model for Republican and Democrat tweets separately, I assign each tweet by the Republicans

[12] This is only for purposes of illustration and the actual models are trained on weekly data, after picking the optimal number of topics using coherence score.

[13] I use the mallet wrapper to run the LDA model because it is considered to be a faster implementation of the LDA model, than the traditional gensim library.

and Democrats to one of the topics for that week. For example, say if a particular week had four optimal topics based on the coherence score, all tweets in that week are assigned to one of the four topics. I use the topic with the highest percentage in a tweet, to assign that topic to that particular tweet. After doing this, I analyze the content in the tweets by the following two methods.

1.5.2.1 Computation of Distance Metrics

I compute similarity and dissimilarity measures between the two topic probability distributions obtained after running the trained LDA model (trained on the pooled tweets of Republicans and Democrats) separately on Democrat and Republican tweets. There are three measures in the literature which seem to serve our purpose. The Hellinger distance is the analogue of measuring the Euclidean Distance between two probability distributions and is a symmetric distance measure.[14]

The Kullback–Leibler divergence also known as relative entropy also measures the distance between two probability distributions, but is not symmetric like the Hellinger distance.[15]

The Jaccard index, also known as the Intersection over Union or the Jaccard similarity coefficient is a measure of the overlap between two sets. The Jaccard index measures the similarity between finite sample sets and is defined as the size of the

[14] The formula for the Hellinger distance between two probability distributions, P and Q is as follows

$$h(P,Q) = \frac{1}{2}||\sqrt{P} - \sqrt{Q}||^2. \tag{1.5}$$

[15] For two probability distributions, P and Q, the Kullback-Leibler divergence is as follows.

$$D_{KL}(P||Q) = \int_{-\infty}^{+\infty} p(x) log \frac{p(x)}{q(x} dx, \tag{1.6}$$

where p(x) and q(x) are the density functions for P and Q respectively.

intersection divided by the size of the sample sets' union. The Jaccard distance is the complement to the Jaccard index. It measures the dissimilarity between two sets and is obtained by subtracting the Jaccard index from 1.[16] I use it to measure the distance between the word distributions used by Republicans and Democrats.

The distances are computed from November 7, 2017, to November 5, 2019, at weekly intervals for 104 weeks. Fig 1.9 shows the pattern of these metrics over the election cycle. The patterns in the distance between the topic distributions measured by the Hellinger distance and Kullback Leibler Divergence are very similar. Hellinger denotes the Hellinger distance whereas the Kullback Leibler divergence is denoted as KLD in the figure. Since the KLD is not symmetric, I calculate the KLD from Democrats to Republicans denoted as KLD_DR and vice versa. The divergence between topic distributions as measured by all these three measures increases as we approach the election. This means that as the election approaches, Republicans and Democrats talk about different topics. The Jaccard distance is calculated between the raw text (after pre-processing the data) used by Democrats and Republicans. I use the vector of all words together used by Democrats and Republicans, and the Jaccard distance, in this case, is denoted by JD_DR. I also use vectors containing a list of words, each list being one tweet and the Jaccard distance, in this case, is denoted by JD_DLRL. According to Fig 1.9 the distance between the word distributions fall as we approach the election or Republicans and Democrats use similar words as we approach the election but not to a very high degree.

[16] The Jaccard distance is defined as follows:

$$d_J(A,B) = 1 - J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}, \tag{1.7}$$

where A and B are finite samples.

1.5.2.2 Euclidean Distance for the dominant topic

The LDA model assigns each tweet to multiple topics. For the next part of the analysis, I find the dominant topic for each tweet and assign it to that particular topic. I calculate the Euclidean Distance between the vectors of the fraction of tweets devoted to each topic by the Democrats and Republicans each week. For example, if there are s topics in a particular week, I obtain two vectors D_s and R_s for that week. Both these vectors have s elements. Let the first element of D_s vector be denoted by $D_s(1)$. Then,

$$D_s(1) = \frac{n(Tweets\, by\, Democrats\, which\, belong\, to\, topic\, 1)}{n(Total\, tweets\, by\, Democrats)}, \tag{1.8}$$

where n denotes the cardinal number. I similarly compute all the elements of D_s and R_s, and find the Euclidean Distance between these two vectors. This is denoted by $Score$.

$$Score \quad = \quad d(D_s, R_s) \quad = \quad \sqrt{\begin{aligned}(D_s(1) - R_s(1))^2 + (D_s(2) - R_s(2))^2 + ... \\ + (D_s(s) - R_s(s))^2\end{aligned}} \tag{1.9}$$

$Score$ helps in understanding the between topic variability in the dominant topics used by Democrats and Republicans. A low value of $Score$ implies that both the parties devote similar weights to the various dominant topics they use in their tweets, whereas a higher value of $Score$ implies that they talk about different topics. Fig 1.10 shows the pattern that the $Score$ metric follows over the election cycle. As the election approaches, there is a rise in the value of the score, which again suggests that Democrats and Republicans are talking about different topics in the approach to the election.

One might wonder if the between topic variability is increasing just due to the number of topics increasing in the approach to the election. However, Fig 1.11 shows that the number of topics in a week decrease in the approach to the election. Therefore, despite the absolute number of topics decreasing, the between topic variability increases corroborating the fact that Democrats and Republicans do devote their tweets to different topics as the election approaches.

1.5.2.3 Euclidean Distance Interacted with Sentiments

The Euclidean Distance helps us understand the pattern in the usage of topics by Republicans and Democrats but provide no insight into how these topics are being used. To make more sense of the intent of the content used for each topic, I perform sentiment augmented content analysis and compute two types of metrics where sentiment is interacted with the topic. These measures give us a low value if Democrats and Republicans talk about the same topic with similar sentiments, and gives us a high value if they talk about the same topic using different sentiments. The two approaches that I take are as follows.

For the first type of metric, I compute the fraction of positive, negative and neutral shares for each topic for Democrats and Republicans separately, then compute the euclidean distance for each topic between Republicans and Democrats, and add the distances for all the topics. For example, if there are s topics in a particular week, I compute 6 vectors $Positive_{D_s}$, $Negative_{D_s}$, $Neutral_{D_s}$, $Positive_{R_s}$, $Negative_{R_s}$, $Neutral_{R_s}$ with s elements each. The first element for $Positive_{D_s}$, denoted by

$Positive_{D_s}(1)$ is defined as follows,

$$Positive_{D_s}(1) \;=\; \frac{n\begin{pmatrix} Tweets\ by\ Democrats\ which\ belong\ to\ topic\ 1 \\ \hfill and\ have\ a\ positive\ sentiment \end{pmatrix}}{n(Total\ tweets\ by\ Democrats\ which\ belong\ to\ topic\ 1)}, \tag{1.10}$$

where n denotes the cardinal number. I similarly compute all the other vectors. My first measure Sum_Frac_Dis, is defined as follows,

$$Sum\_Frac\_Dis = \frac{\begin{matrix} d(Positive_{D_s}, Positive_{R_s}) + d(Negative_{D_s}, Negative_{R_S}) \\ \hfill + d(Neutral_{D_s}, Neutral_{R_s}) \end{matrix}}{}, \tag{1.11}$$

where $d(.,.)$ denotes the Euclidean distance between two vectors.

Second, instead of computing the fractions of positive tweets for a particular topic for each group, I use the intensity of the sentiment to derive the metric for Euclidean distance. Therefore, instead of counting the number of positive, negative or neutral tweets for each group, I compute the intensity of positivity, negativity and neutrality in the tweets. For s topics, I again compute the six vectors $Positive_{D_s}$, $Negative_{D_s}$, $Neutral_{D_s}$, $Positive_{R_s}$, $Negative_{R_s}$, $Neutral_{R_s}$. The first element for $Positive_{D_s}$, denoted by $Positive_{D_s}(1)$ is defined as follows,

$$Positive_{D_s}(1) = \begin{matrix} Mean\ value\ of\ positive\ score\ for\ tweets\ by\ Democrats \\ which\ had\ a\ positive\ sentiment\ and\ belonged\ to\ topic\ 1 \end{matrix} \tag{1.12}$$

The metric Sum_Dis is then calculated as follows,

$$Sum\_Dis = \frac{\begin{matrix} d(Positive_{D_s}, Positive_{R_s}) + d(Negative_{D_s}, Negative_{R_s}) \\ \hfill + d(Neutral_{D_s}, Neutral_{R_s}) \end{matrix}}{}, \tag{1.13}$$

where $d(.,.)$ denotes the Euclidean distance between two vectors.

Figure 1.12 shows the pattern of Euclidean Distance interacted with the sentiment for the fraction of positive, negative and neutral tweets of a particular topic. Unlike the *Score* metric discussed in the last subsection, the value of the *Sum_Frac_Dis* falls as we approach the election. This suggests that Democrats and Republicans use similar sentiment to talk about common topics as they approach the election. We get the same trends when we use the intensity of the positive, negative or neutral sentiment instead of the fractions. The figure is shown in the Appendix.

1.5.3 Bayesian Ideal Point Estimation

The third way I try to capture the short-run polarization is by computation of the politicians' ideological estimates using their tweets by the method of Bayesian ideal point estimation. This method developed by Eady (2018) uses the URL sharing behaviour to infer the politicians' ideology.[17] The idea is that politicians will share more URLs from a media source close to them in the ideological spectrum. A Monte Carlo simulation is then performed to infer the ideologies of the politicians.

To estimate the ideologies of the politicians, I extract the URLs from the tweets of the Representatives. The extracted URLs are shortened URLs, and cannot be directly used. The URLs are then expanded into their long-form by querying the server using the shortened URLs. I then extract the URLs' domain names and compute an adjacency matrix of how many times each Congressperson has tweeted any particular website. Retweets are included in this analysis, as retweets also signify the reiteration of the original tweet's content by the person who is retweeting the original tweet. I remove all social media domain names from the adjacency matrix such as google.com,

[17] They talk about the analysis in 'Trying to understand how Jeff Flake is leaning? We analyzed his Twitter feed — and were surprised" in The Washington Post, October 5,2018.

facebook.com, instagram.com and others as these do not have any ideological content of their own.[18] For data sanity purposes, I take only 90 percent of the total shares of news domains, as this helps the data to be devoid of obscure websites, which had only been mentioned once or twice, and other obscure names, which show up in the adjacency matrix due to technicalities of domain name extraction. After this, I employ the empirical strategy used by Eady (2018)[19]. I differ from Eady's strategy in that I do not specify whether an individual is a Republican and Democrat, and only use the information from the URL sharing behaviour to get my estimates.

The empirical strategy is implemented as follows:

$$y_{im} \sim \text{NegBin}(\alpha_{im}, \psi_i \psi_m) \tag{1.14}$$

$$\alpha_{im} = \exp(\theta_i + \lambda_m - ||\tau_i - \upsilon_m||^2), \tag{1.15}$$

where y_{im} denotes the count of shares of domain m shared by user i, in our case a member of the House of Representatives; θ_i denotes a user-specific intercept which essentially means that some Congresspeople are more active on Twitter and may indulge in higher URL sharing activity; λ_m denotes a domain-specific intercept which similarly accounts for the fact that some domains might have a higher probability of being shared than others, and ψ_i and ψ_m denote user-specific and domain-specific dispersion parameters respectively to capture the predictability in the model. The quantities of interest are denoted by τ_i, which represents the ideology of user i, and

[18] I also remove any website that has "house.gov" in its URL, because the Representatives seem to be using Twitter as a platform to broadcast those websites.

[19] He has developed a mediascores package in R available on Github for faster and more straightforward implementation of the Bayesian Ideal Point Estimation, which I have used for my analysis.

υ_m, the ideology of m, the website being shared. As the term $-||\tau_i - \upsilon_m||^2$ makes clear, the larger the ideological spatial distance between the ideology τ_i of the user and the ideology υ_m of the website, the less likely the user is to share stories from that website. Priors are placed on the model parameters as follows:

$$\theta_i \sim \text{Normal}(\mu_\theta, \sigma_\theta) \tag{1.16}$$

$$\lambda_m \sim \text{Normal}(0, \sigma_\lambda), \tag{1.17}$$

where uniform prior distributions are placed on the hyper-parameters μ_θ, σ_θ and σ_γ. The variance parameters ψ_i and ψ_m, are given common distributions $\psi_i \sim \text{InvGamma}(\psi_a^{(i)}, \psi_b^{(i)})$ and $\psi_m \sim \text{InvGamma}(\psi_a^{(m)}, \psi_b^{(m)})$, with uniform priors on the hyper-parameters. For identification, the parameters representing news media ideology are centered on 0: $\upsilon_m \sim \text{Normal}(0, \sigma_\upsilon)$. Lastly, the model's direction needs to be set, such that high values represent ideological liberalism or conservatism. The researcher fixes the anchors (e.g. nytimes.com, foxnews.com), such that the ideology of the first media organization defines the low end of the scale (in this example, liberal), and the second, the high end (in this example, conservative).

In my case, I choose the anchors to be domains shared by one of the parties more relative to the other party. Using this methodology, I compute the number of times a Democrat has shared a particular domain divided by the number of times it has been tweeted totally, to be the most Democratic domain and similarly for Republicans. I call this the most differential domain for Democrats and Republicans respectively. I also eliminate the domains shared by less than 2 percent of the respective group of politicians. They would be fringe websites, and will not be efficient for the Monte

Carlo Simulation. One of the upshots of this analysis is that I choose the anchor websites dynamically and objectively based on the politicians' tweeting behaviour and not subjectively as has been done in the original implementation of the model. The most differential domain by Democrats moves between vox.com, nytimes.com, cnn.com, npr.org and others. The most differential domain by the Republicans lingers between foxnews.com, wsj.com, www.washingtonexaminer and others. I had also replicated the analysis with the most highly tweeted domain, and in that case, the Democrat anchor website is overwhelmingly nytimes.com. Since nytimes.com is not considered to be the most extreme left leaning news media organization, I think fixing the anchor website to be the most differentially highly tweeted domain makes more sense. Fig 1.13 and Fig 1.14 shows the most popular domain names among the Democrat and Republican politicians respectively. Another point of departure from the traditional model as implemented is that I do not assign separate groups to Democrats and Republicans as I want only the URL sharing behaviour to inform their ideological points, and not to get biased by their group identity, as already explained before. Therefore, the estimates in my model are not biased by ex-ante group identity.

I also divide the politicians depending upon whether they are contesting in a competitive or non-competitive race. As mentioned in Section 1.4, solid districts are considered as non-competitive districts, whereas Likely, Lean or Toss-up districts are considered as competitive districts. Each week the politician contesting from a particular district is considered to be competitive if that particular district is competitive in that week and to be non-competitive if the particular district is non-competitive that week Figure 1.15 shows the mean ideological trend for the Democrats and the Republicans calculated using Bayesian Ideal Point Estimation. *Rep_Comp* refers to

Republicans in competitive districts whereas $Rep\_Non-Comp$ refers to Republicans in non-competitive districts and similarly for Democrats. The figure shows that the ideological score increases as we approach the election both for competitive and non-competitive districts. Figure 1.16 shows the ideological polarization between Republicans and Democrats a year before and after the election. The ideological difference is calculated as the difference between the mean ideological score of Republicans, and Democrats. ID_Comp shows the difference in the ideological difference between in the competitive districts and $ID\_Non-comp$ shows the ideological difference in the non-competitive districts. The ideological difference decreases as we approach the election both for competitive and non-competitive districts.

1.5.4 Mentions Network Analysis

For the fourth part of the ideological metric, I calculate the polarization in the mentions network in my data set. This gives us an idea about the affective polarization within the network. Figure 1.17 visually shows the polarization in the mention network. As we can see, the network is heavily polarized in terms of the interactions members of the two parties have with one another. Since I want to look at the change in polarization in a time series fashion, traditional network analysis measures such as nodes, degree or centrality do not help.

To compute the degree of affective polarization in this network, I compute the shares of how many times Democrats mention other Republicans over 104 weeks, and vice-versa[20]. Figures 1.18a and 1.18b show that both parties mention each other

[20] To make the analysis more accurate, I collect the handles of all Democrats and Republicans for the present House and Senate, and also for the previous House, as well as the current and past President. Therefore, the Republicans whom Democrats mention are Republicans in the current House of Congress and include Senators and House members in the 115th House of Congress, with a verified official handle

more negatively with the approach of the election. However, while there is a drop in negative mentions of Republicans by Democrats, there is a significant increase in negative mentions of Democrats by Republicans after the election. Both the estimates, however, keep falling as we move away from the election. It is important to note that these are broad level trends, and some of these trends change direction when we zoom in to the last eight weeks of the election. The results of what happens in the last eight weeks are discussed in Section 1.6.

1.5.5 Retweet Network Analysis

I compute similar measures of polarization for the retweet network as well. Figures 1.19a and 1.19b show that the negative retweets of out-group increases as we approach the election. There is a slight drop in negative retweets of Republicans by Democrats and a slight increase in negative retweets of Democrats by Republicans and negative retweets of the out-group decreases as we move away from the election.

1.6 Empirical Analysis

1.6.1 Regression Discontinuity Analysis

After computing the metrics of ideological polarization, I use the non-parametric Regression Discontinuity Design to see if there is a significant discontinuous jump before and after the election. I use weeks to the election as my running variable, with the cut-off at 0 and the metrics that I have already computed as my outcome variables. The RDD set-up works very well in this scenario. Although the election is not an exogenous event as is generally the requirement for an RDD, this works in our favour because we are trying to measure the effect on ideology as soon as the anticipation of an impending election goes away. For this reason, I use a non-parametric RDD to look at only a narrow window before and after the election. A data-driven approach is used to find the right bandwidths for the regression as outlined in Calonico et al. (2014).[21] I perform non-parametric Regression Discontinuity in Time as performed in Davis (2008). The results for varying degrees of polynomials used are reported. The non-parametric RDD in time is estimated using the following equation.

$$y = f(t) + \epsilon, \tag{1.18}$$

where y denotes the various metrics that I have calculated and t denotes time in weeks, and is the running variable.

1.6.1.1 Results

In the RDD estimates, I only find a significant increase in the number of common hashtags used which shows that Republicans and Democrats start talking about similar things at a higher rate just after election compared to that before the election as shown in Table (1.2). None of the other estimates have any significant discontinuities as shown in Tables 1.3 - 1.7. There are some effects in the mention network analysis and retweet network analysis. $Mentioned\_Dem\_Negatively$ estimates by Republicans increase right after the election, as shown in Table 1.6. In the retweet network analysis, estimates of $Retweeted\_Rep\_Negatively$ by Republicans decreased right after the election, as shown in Table 1.7.

[21] I use the rdrobust package in R for the Regression Discontinuity estimation. The package uses a data-driven methodology to select the best bandwidth

1.6.2 OLS Estimates for sub sample

It is clear by looking at the smoothed lines in the graphs of the metrics that they vary considerably depending on where one is in the election cycle. The locally polynomial regression lines shown in the graphs help us understand how these metrics vary. Therefore, the metrics give us a sliding window view of what is happening to polarization at any point in time. As is evident, polarization is different depending upon where one is in the election cycle, and does not have a long term trend. Therefore, it does not make much sense to infer the effect of the election by using the entire time series. Another potential challenge in inferring any causality from these graphs is that many events can influence politicians' tweeting behaviour, such as significant worldwide events or primaries. There could also be potential seasonality effects in the time series. Therefore, to understand what is happening just before the election, one needs to focus on a narrow window close to the election.

As a specific case for illustrative purposes, I look at a window of 8 weeks before and after the election. I choose eight weeks because there are no primaries contested in the last eight weeks for the federal election. There is also no major worldwide event to influence politicians' tweeting behaviour. Therefore, it is reasonable to assume that all tweets in the last eight weeks of the election cycle will be about the midterm election, and this helps us get closer to the effect of the election.

I run a simple linear regression model to estimate the slopes of the ideological metrics that I have computed before and after the elections. It is important to note that I am not trying to show causality or compute exact estimates about the magnitude of change, but I am more interested in the direction of change. The direction of change sheds light on whether polarization decreases or increases as we approach the election. The equation that I estimate is

$$y_t = \alpha + \beta Week + \epsilon, \tag{1.19}$$

where y_t represents the various metrics that I computed. I estimate two separate slope coefficients one before the election and one after the election. For the pre-election version, $Week$ increases from 1 to 8 as we get closer to the election, with the 8^{th} week being closest to the election. In the post-election version, $Week$ increases as we get further away from the election. The 8^{th} week is the furthest away from the election. Another way to think about this would be that $Week$ increases in the positive direction of the time for the pre-election and post-election specification.

1.6.2.1 Results

The slopes for the hashtag estimates are negative in the last eight weeks of the election, as shown in Table 1.8. This implies that Democrats and Republicans decrease their use of similar hashtags as they approach the election conditional on the top hashtags. The estimates for $Hashtag_{50}$ and $Hashtag_{100}$ are significant at 5 percent level. The estimates also decrease as the politicians move away from the election. However, there is a significant increase in the number of common hashtags just after the election, as shown in Table 1.2. The slopes of the $Inverse\_Score\_std$ metrics are also negative, implying inverse score falls and therefore distance increases as we approach the election. This means that as elections get closer, conditional on using the same hashtags politicians from different parties use increasingly different sentiments to talk about those hashtags. The slope of the measure is negative after the election too, but there is an increase in the inverse score as soon as the election is over as shown in Table 1.2. These measures hint at an increase in polarization or movement away from the median voter in anticipation of the election.

Similarly, the slopes of Hellinger distance and the Kullback-Leibler divergence are positive as shown in Table 1.9 implying an increase in topic divergence. Even when we focus only on the dominant topics and use the distance between the relative shares that Democrats and Republicans devote to such topics, the slope is positive implying that distance increases as we approach the election as shown in Table 1.10. The combined evidence suggests that Democrats and Republicans increasingly talk about different agendas with the approach of the election. Therefore, both in the case of hashtags, which are very context-specific framing devices and broad topics, Democrats and Republicans grow increasingly divergent with the approach of the election.

However, the results are opposite when we consider the words used in the tweets and when we augment the topic modelling with sentiment analysis. The Jaccard Distance slope between word distributions is negative, as shown in Table 1.9. This implies that politicians increasingly use similar words in their tweets. When the topic analysis is augmented with sentiment scores, the Euclidean Distance score decreases which means that politicians increasingly use the same sentiment to talk about different topics as shown in Table 1.10. Therefore, as implemented through sentiment augmented topic modelling, the content analysis hints at a decrease in polarization and movement towards the median voter in anticipation of the election.

The ideological difference between Republicans and Democrats as computed using the Bayesian ideal point estimation also decreases for competitive and non-competitive districts, as shown in Table 1.11. The ideological difference in the competitive districts increases after the election, whereas the non-competitive districts' difference decreases after the election. This also suggests a decrease in polarization and movement towards the median voter as the election approaches.

Negative mentions of Republicans by Democrats increases as we move into the election, falls once the election is over, then decreases for some time and increases again as we move away from the election. Negative mentions of Democrats by Republicans increases as we move into the election, rises (significantly) once the election is over, then decreases for some time and increases again as we move away from the election. Negative retweets of Republicans by Democrats decrease in the last eight weeks of the election. They fall (significantly) once the election is over, and increase in the eight weeks after the election. Negative retweets of Democrats by Republicans increase as we approach the election, decreases after the election and falls after that.

The combined pieces of evidence suggest that while Democrats and Republicans become more polarized in their agenda-setting behaviour with the approach of the election, they become less polarized in terms of the content shared within a particular agenda. Whereas the between-agenda or between-topic variability increases with the approach of the election, the within-topic variability decreases. One way to think about this would be that while politicians are trying to appeal to their extreme electoral bases through their agenda-setting behaviour, they try to appeal to the median voter or the swing voters by remaining more moderate in their content within the diverse agendas. This could also be because while faithful voter bases might be lured by token gestures or the appearance of extremism, more moderate and attentive voters might need more content to win them over.

1.7 Conclusion

In this paper, I collect tweets from incumbent Representatives in the 116^{th} House of Congress at a weekly interval one week before and after the 2018 midterm election. I then use the Twitter data to construct several estimates of political ideology using hashtag analysis, topic modelling, Bayesian Ideal Point Estimation, and analyze the mention and retweet networks of the politicians.

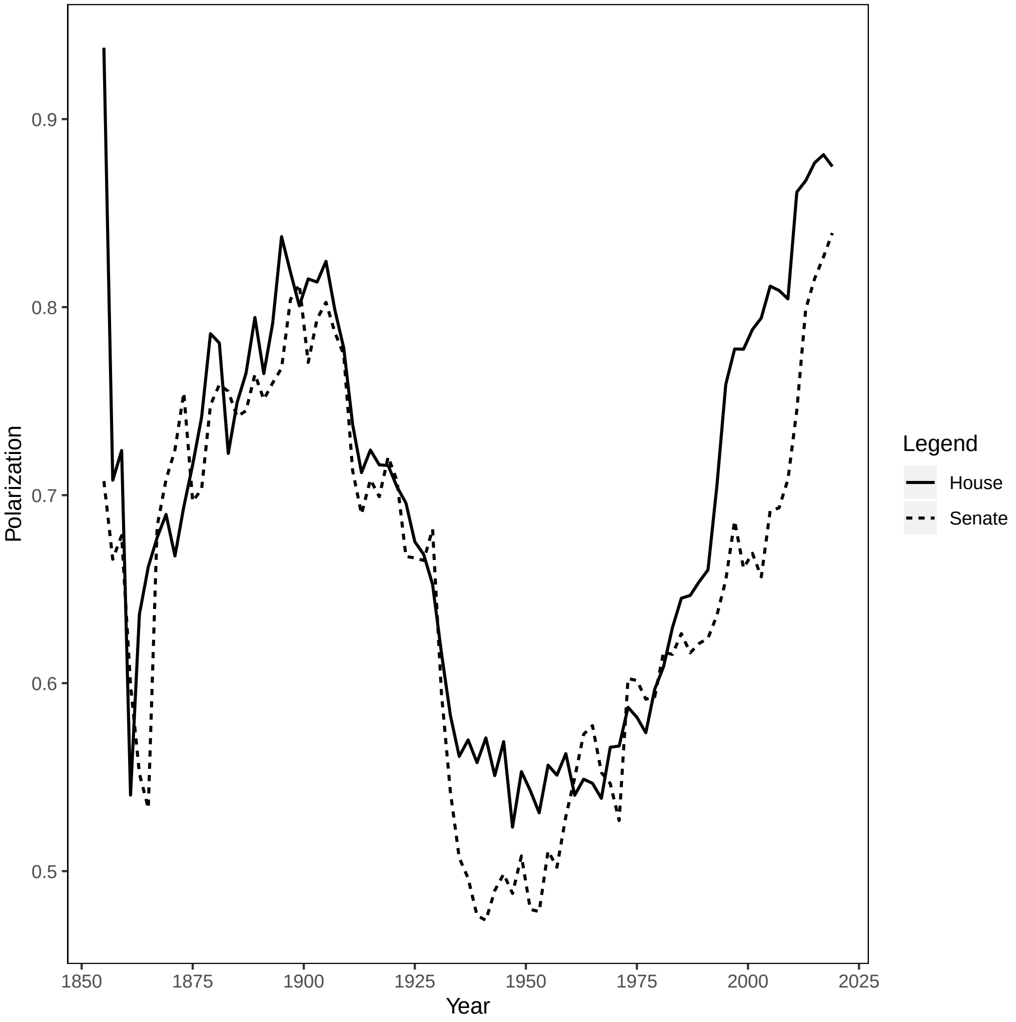
There are two ways one can interpret the result from this paper. From a Frequentist perspective, the statistical insignificance of the Regression Discontinuity estimates and the sub-sample estimates (except for the Hashtag estimates and some of the network estimates) suggest that there is no discernible discontinuity at the election. It also implies no significant change in behaviour at the level of discourse either before or after the election.

However, one could also argue that there is very little variation in the independent variable for the OLS subsample estimates. Therefore, it is difficult to get precise estimates with low standard errors. Adopting a Bayesian perspective helps us infer some patterns from the estimates that can inform our priors which can be later validated/rejected through a future project. When we use the Bayesian perspective, we find some interesting patterns. Polarization, as measured by broad level agenda-setting behaviour such as hashtags or topics, is found to increase with the approach of the election. However, when we shift our measuring instrument to the similarity in words used, sentiment augmented topic analysis or ideological scores inferred from the media sharing activity, we find polarization to decrease with the approach of the election. This suggests that there is convergence in agenda-specific positioning while there is increasing divergence in preferred electoral agenda setting.

A potential future research area would be to repeat this analysis for multiple election periods and check if the patterns that we find here repeat in multiple election periods or if it is something unique to this election period. The methodologies developed in this paper could also be extended to understand voters' behaviour and preferences during an upcoming election and get a more accurate understanding about how voters' preferences change as we get close to the election.

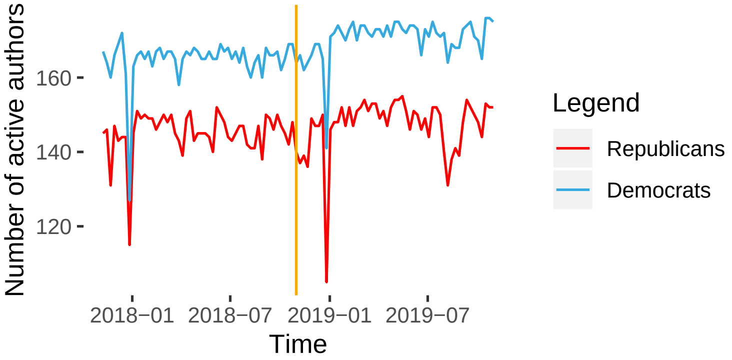
1.8 Figures and Tables

Figure 1.1: Polarization as measured by DW-Nominate scores over the years



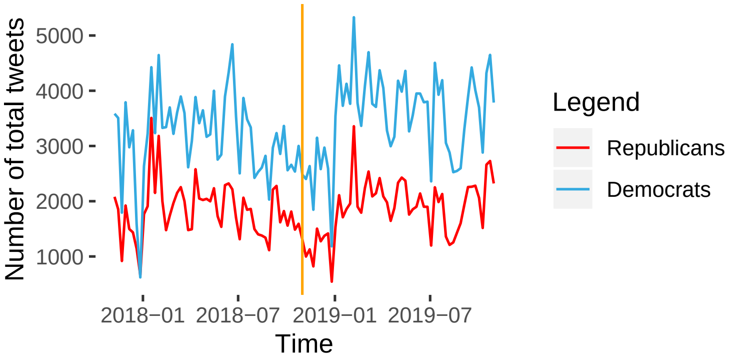
Notes: Difference between mean ideological positions for Republican and Democrat politicians from 1855 to 2019 along dimension 1 of DW-Nominate scores, using data from voteview.com. An almost similar graph is reproduced by the estimates of dimension 1 constructed by Poole and Rosenthal.

Figure 1.2: Total number of active politicians in each week.



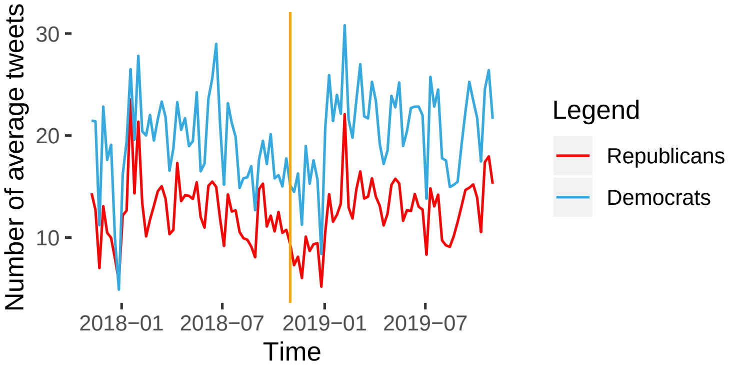
Notes: The number of active authors (authors who have tweeted atleast one tweet in a week) are counted for each week. The orange line indicates the election.

Figure 1.3: Total number of tweets in each week.



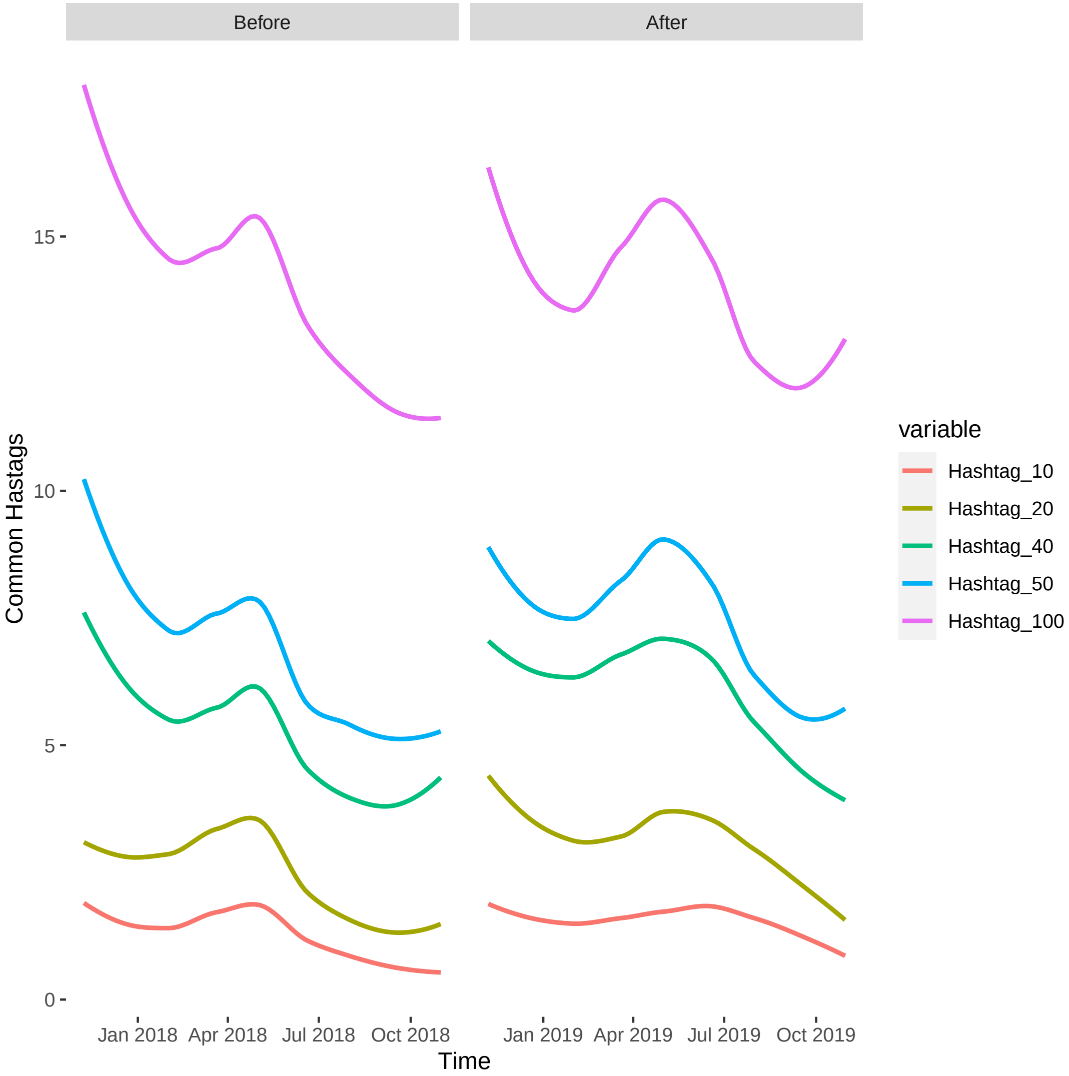
Notes: The total number of tweets are plotted for each week. The orange line indicates the election.

Figure 1.4: Number of average tweets in each week.



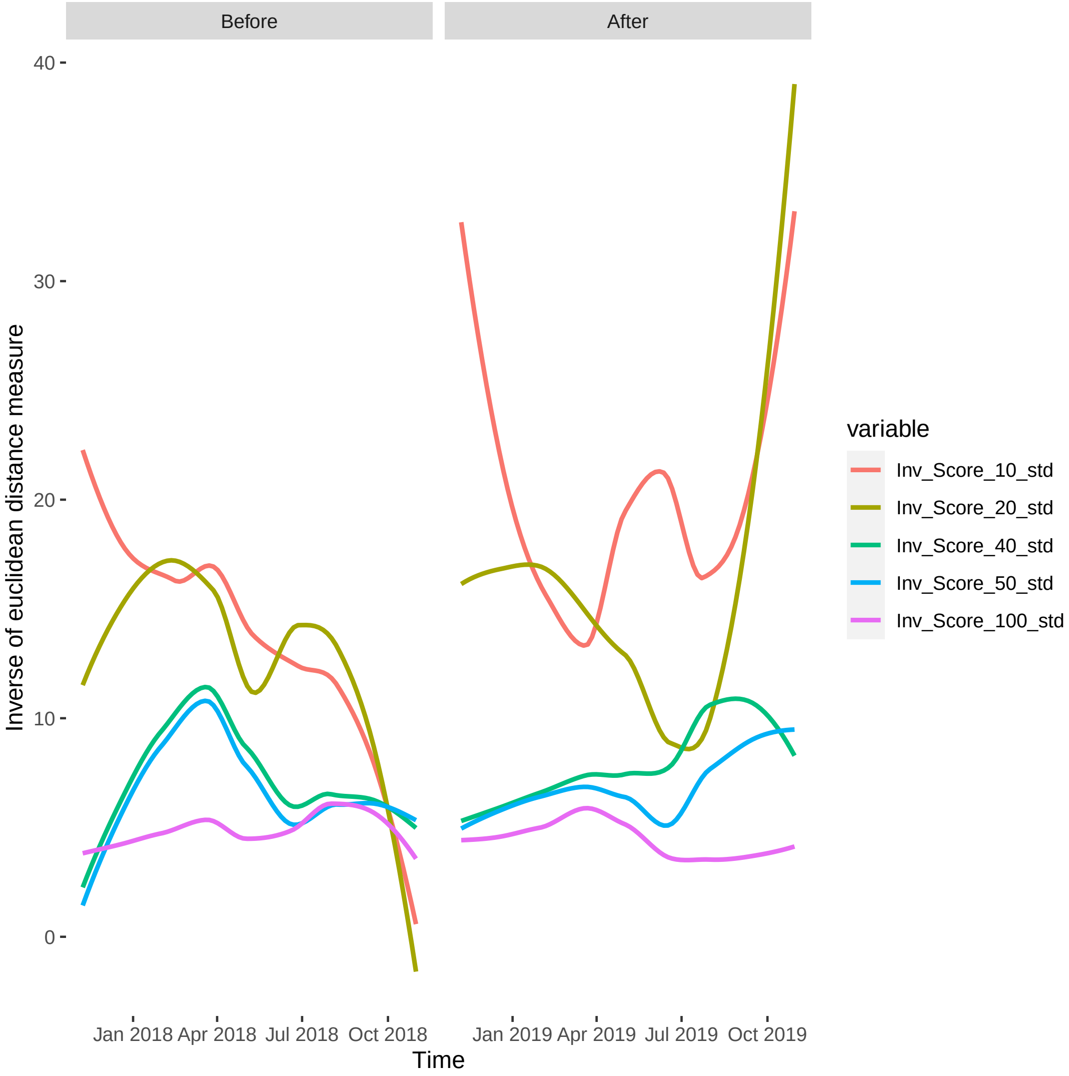
Notes: The number of tweets sent out on average each week. The numbers plotted in this graph is obtained by dividing the total number of tweets by the number of active authors. The orange line indicates the election.

Figure 1.5: Similarity in hashtags over time



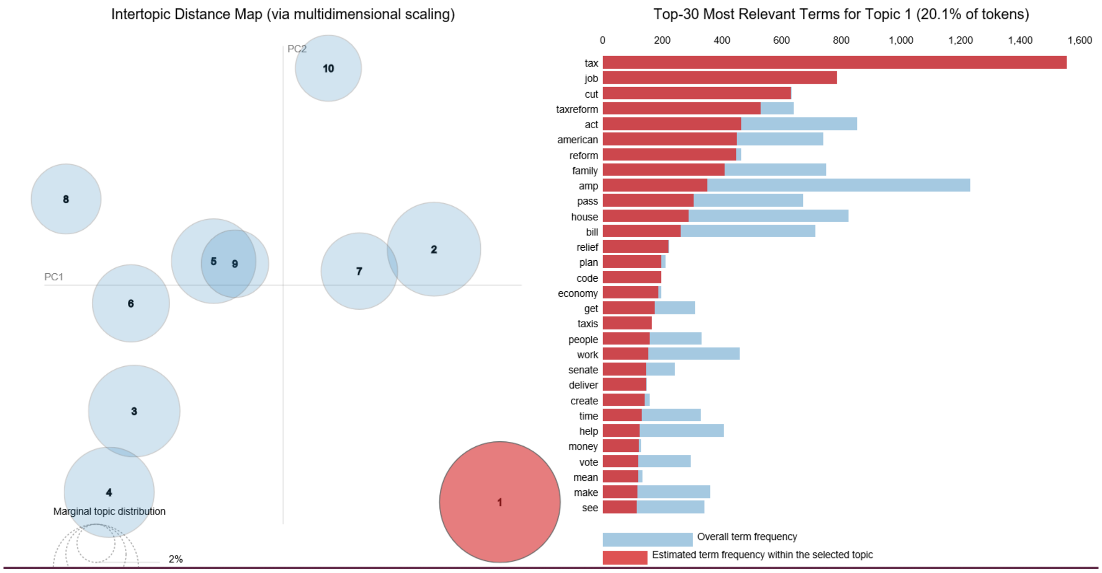
Notes: The figure plots the number of common hashtags used in the top 10, 20, 40, 50 and 100 hashtags used by Republicans and Democrats. The Before facet shows what the trend before the Nov 6^{th} midterm election, whereas the After facet shows the trend after the Nov 6^{th} midterm election.

Figure 1.6: Trend of inverse of standardized Euclidean Distance of hashtags interacted with sentiments



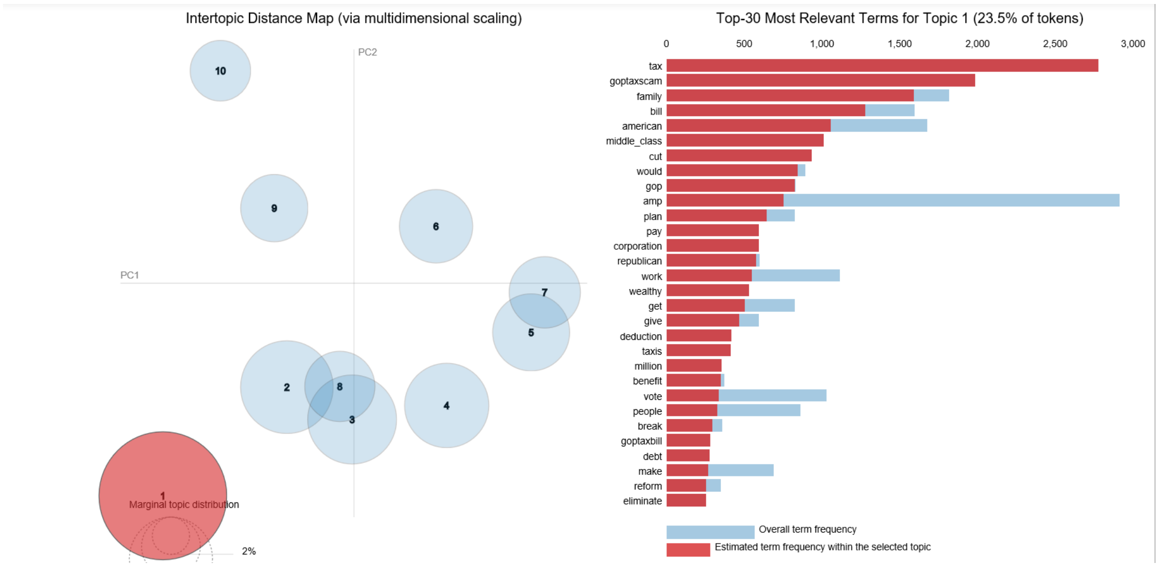
Notes: The figure inverse of the euclidean distance between negative tweets containing hashtags between Republicans and Democrats. A higher value implies low polarization whereas a lower value implies higher polarization. The Before facet shows what the trend before the Nov 6^{th} midterm election, whereas the After facet shows the trend after the Nov 6^{th} midterm election.

Figure 1.7: Token distribution for Topic 1 in Republican tweets



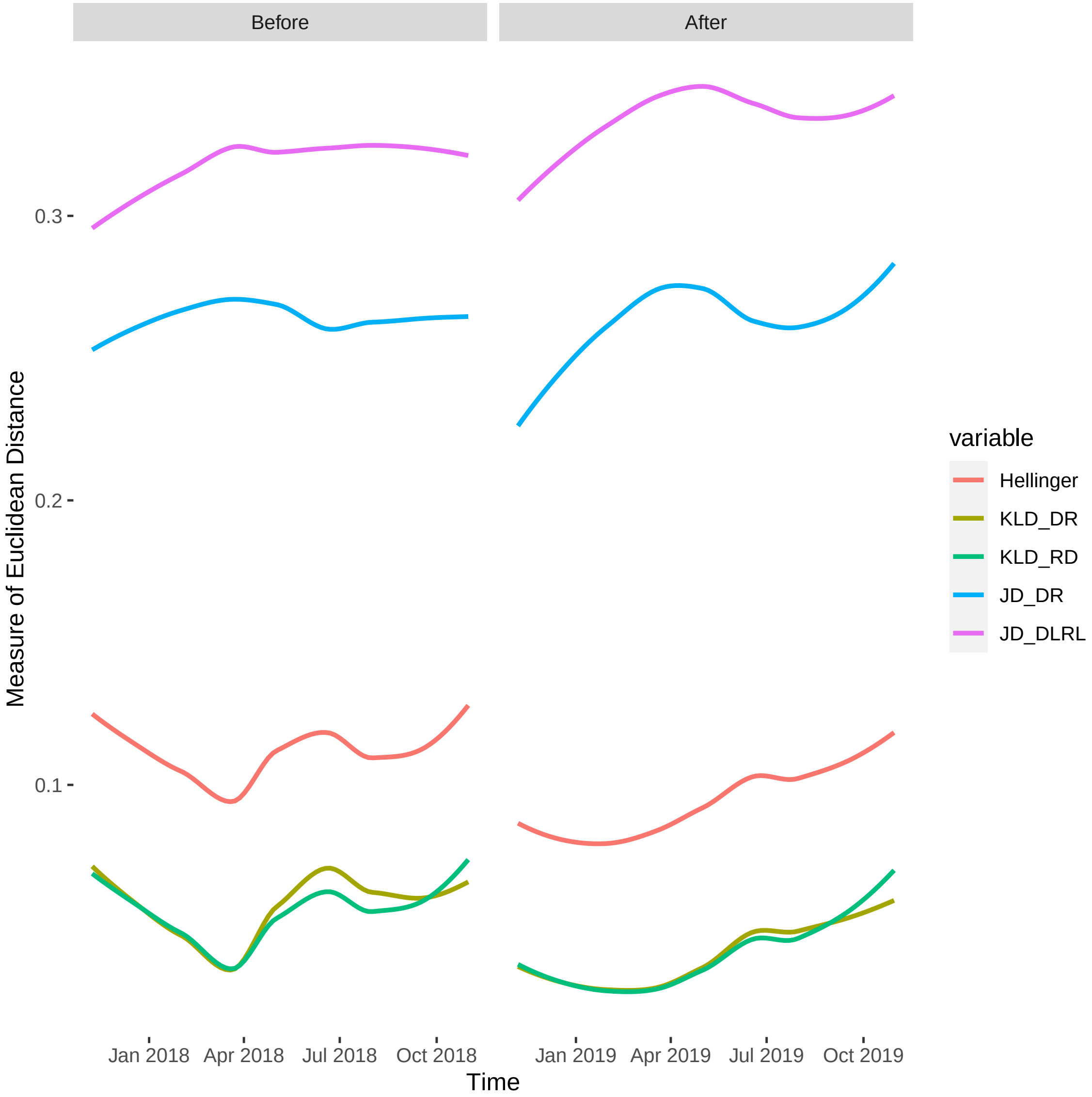
Notes: The top 30 most relevant words used in the first topic in the tweets by Republicans for the month of 7^{th} November- 7^{th} December, 2017 after fitting the LDA model.

Figure 1.8: Token distribution for Topic 1 in Democrat tweets



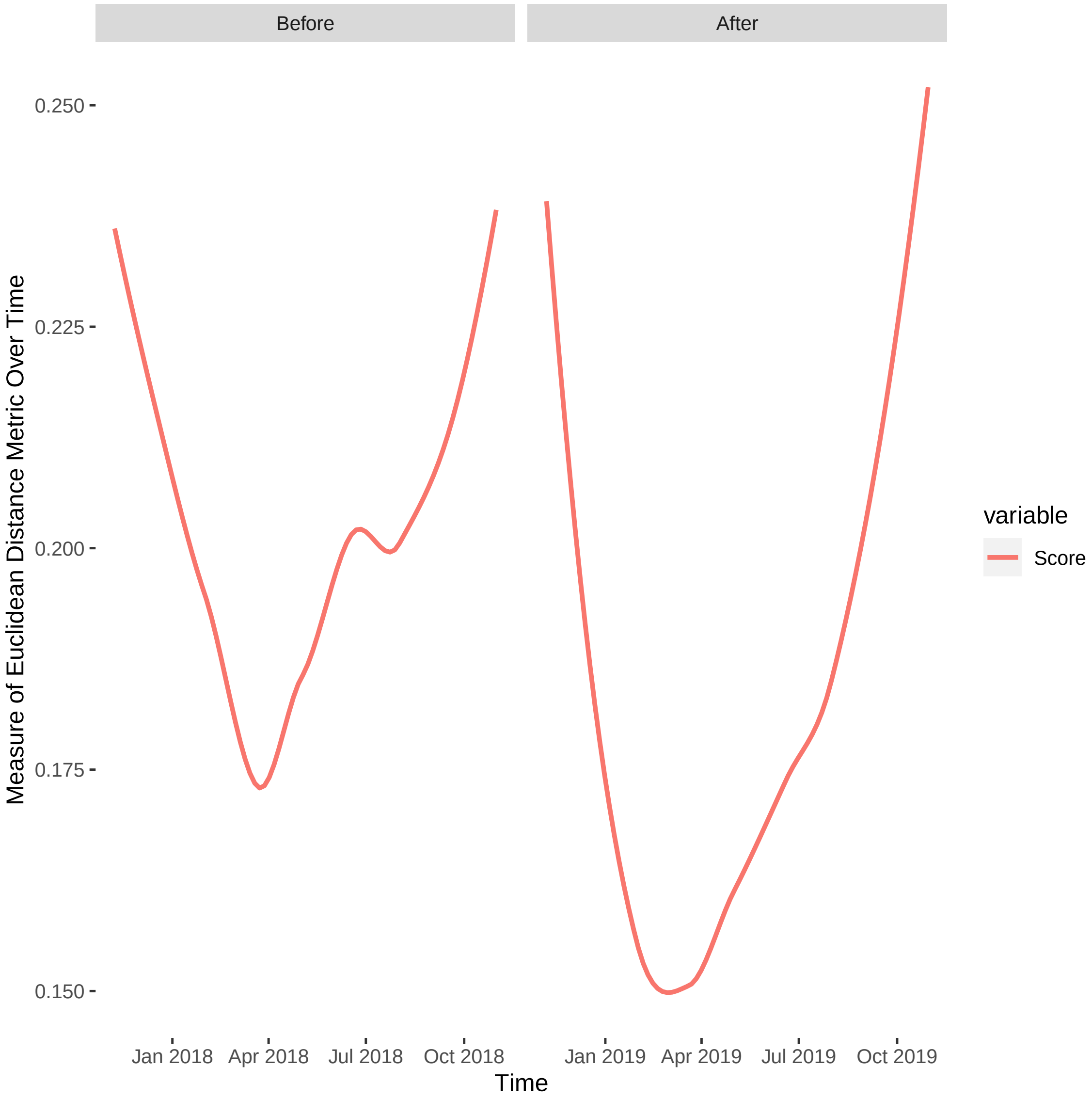
Notes: The top 30 most relevant words used in the first topic in the tweets by Democrats for the month of 7^{th} November- 7^{th} December, 2017 after fitting the LDA model.

Figure 1.9: Distance metrics over time between Democrat and Republicans



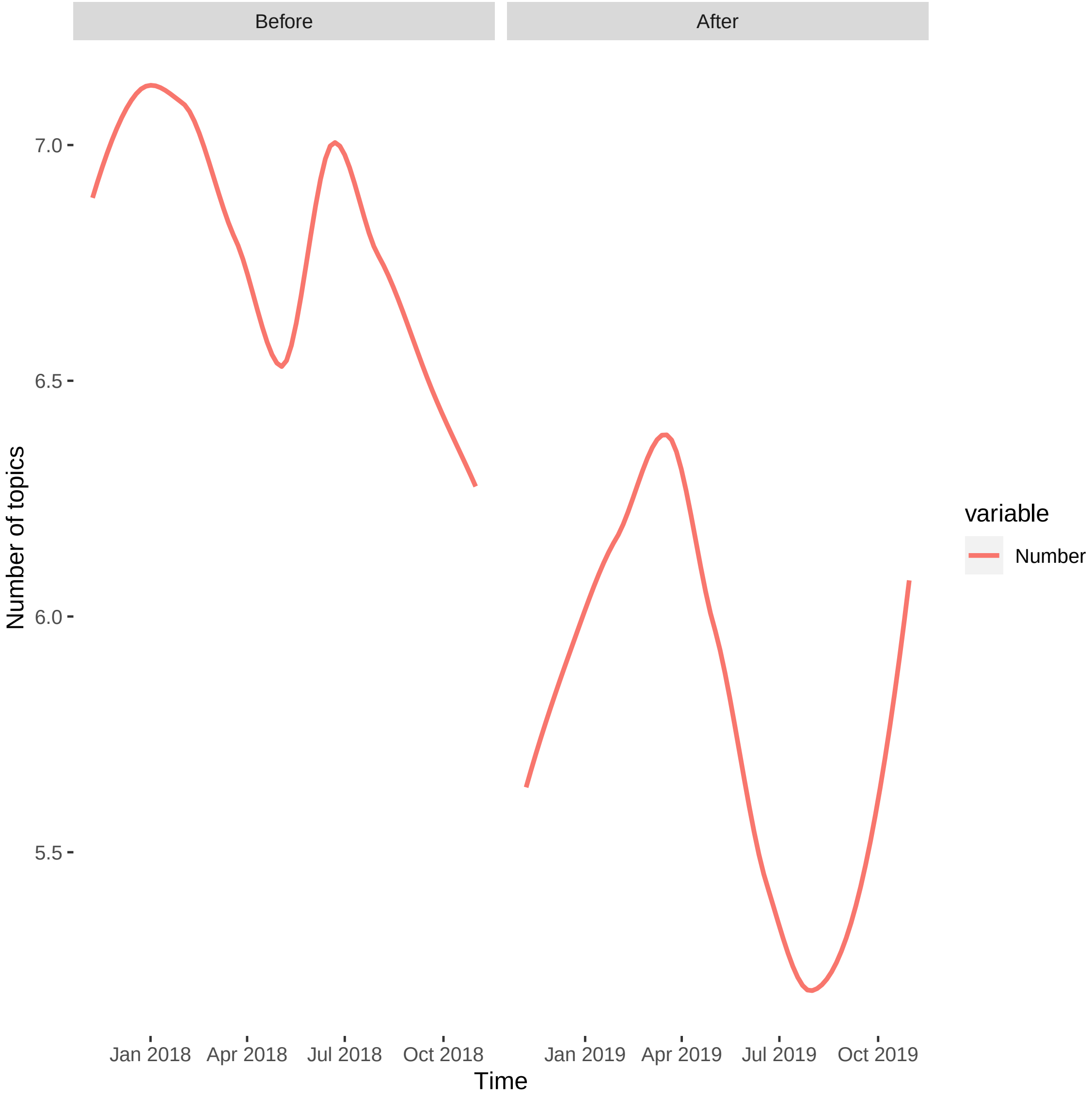
Notes: Euclidean distance between the probability distributions after applying the LDA model as measured by Hellinger, KLD_DR, KLD_RD. KLD refers to the Kullback Leibler Divergence. Distance between the words used and the list of words used as measured by JD_DR and JD_DLRL respectively. JD refers to the Jaccard distance. The Before facet shows what the trend before the Nov 6^{th} midterm election, whereas the After facet shows the trend after the Nov 6^{th} midterm. election.

Figure 1.10: Euclidean Distance between the dominant topics



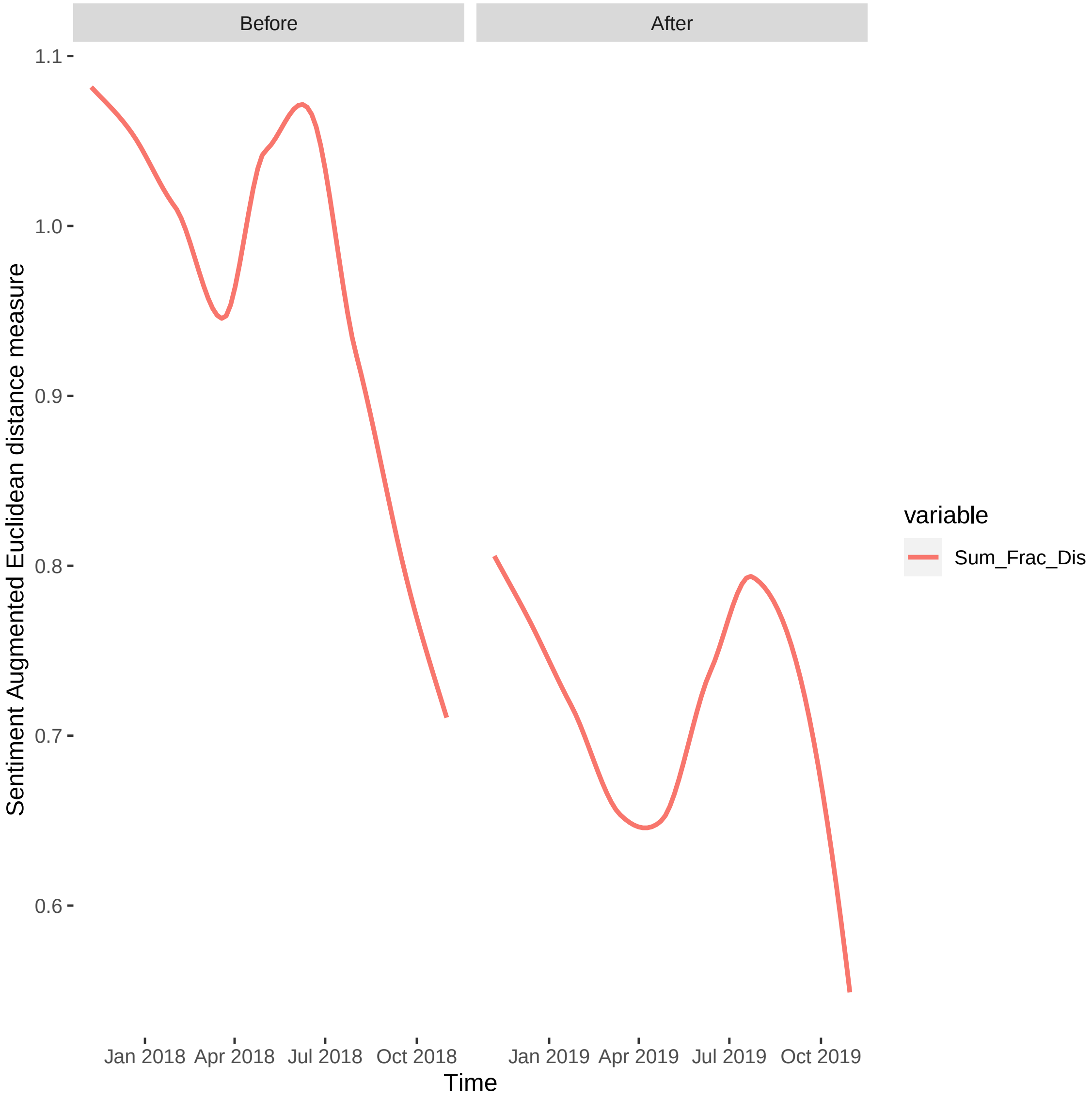
Notes: Score shows the euclidean distance between the vector of fractions of the tweets by Republicans and Democrats devoted to the dominant topic calculate at at a weekly basis. The Before facet shows what the trend before the Nov 6^{th} midterm election, whereas the After facet shows the trend after the Nov 6^{th} midterm.

Figure 1.11: Number of topics in each week according to the optimal LDA model



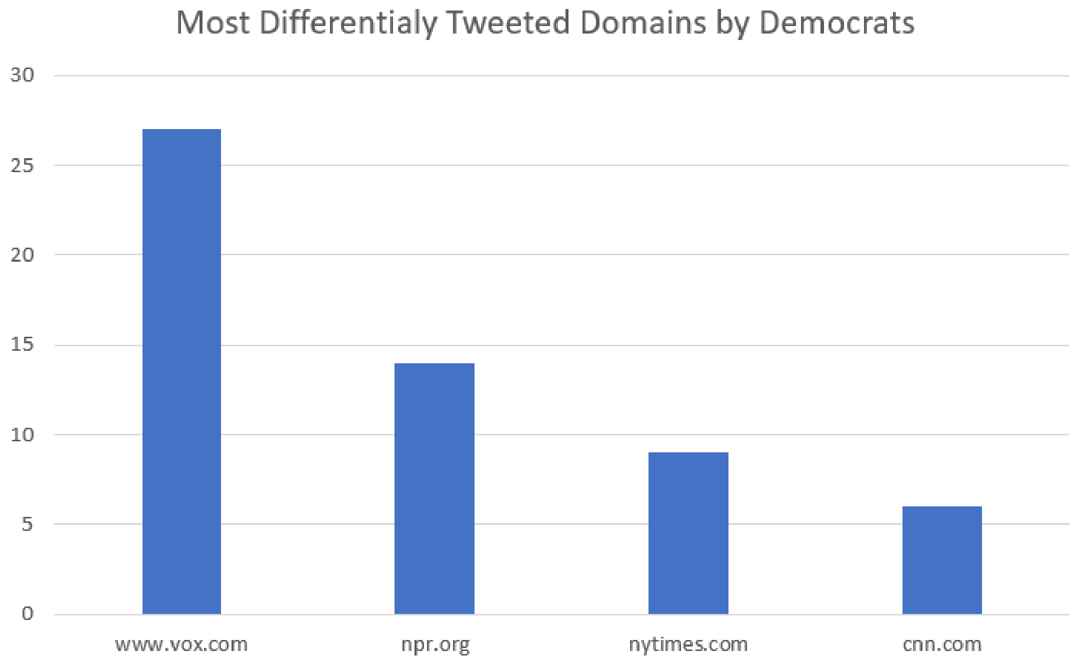
Notes: The graph shows the smoothed curve over the number of topics every week. The number of topics when the coherence scores stops increasing is used. The Before facet shows what the trend before the Nov 6^{th} midterm election, whereas the After facet shows the trend after the Nov 6^{th} midterm.

Figure 1.12: Euclidean Distance for fraction of positive, negative and neutral sentiments in a topic between Democrats and Republicans



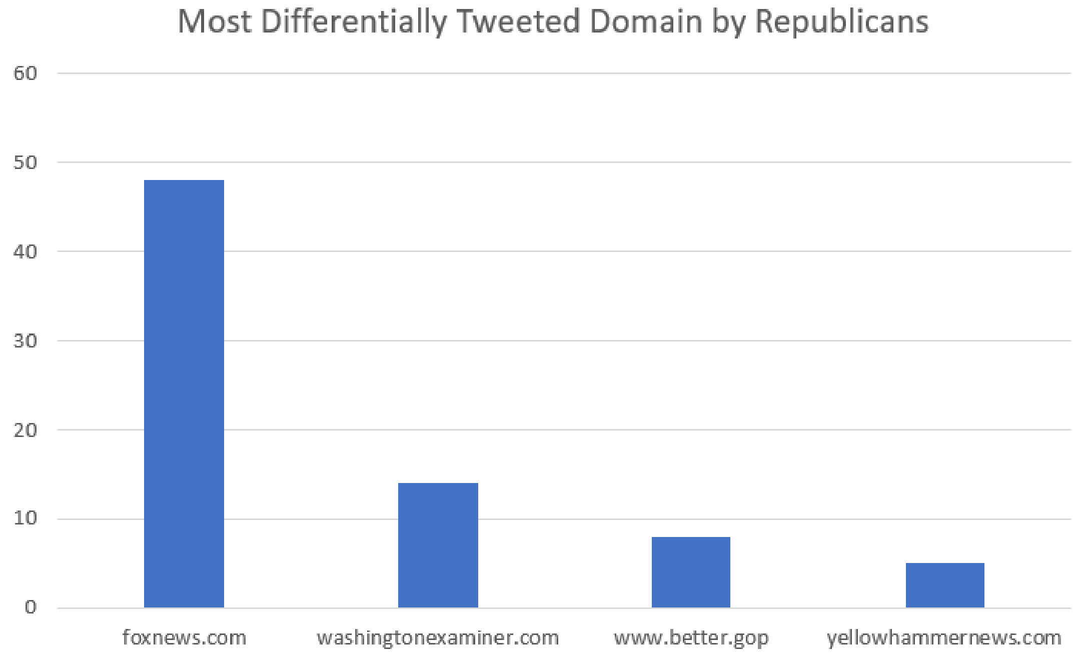
Notes: Sum_Frac_Dis measures the euclidean distance between the vector of fractions of tweets by Democrats and Republicans used with a positive, negative and neutral sentiments. The Before facet shows what the trend before the Nov 6^{th} midterm election, whereas the After facet shows the trend after the Nov 6^{th} midterm.

Figure 1.13: Domains which have tweeted relatively more by Democrats compared to Republicans



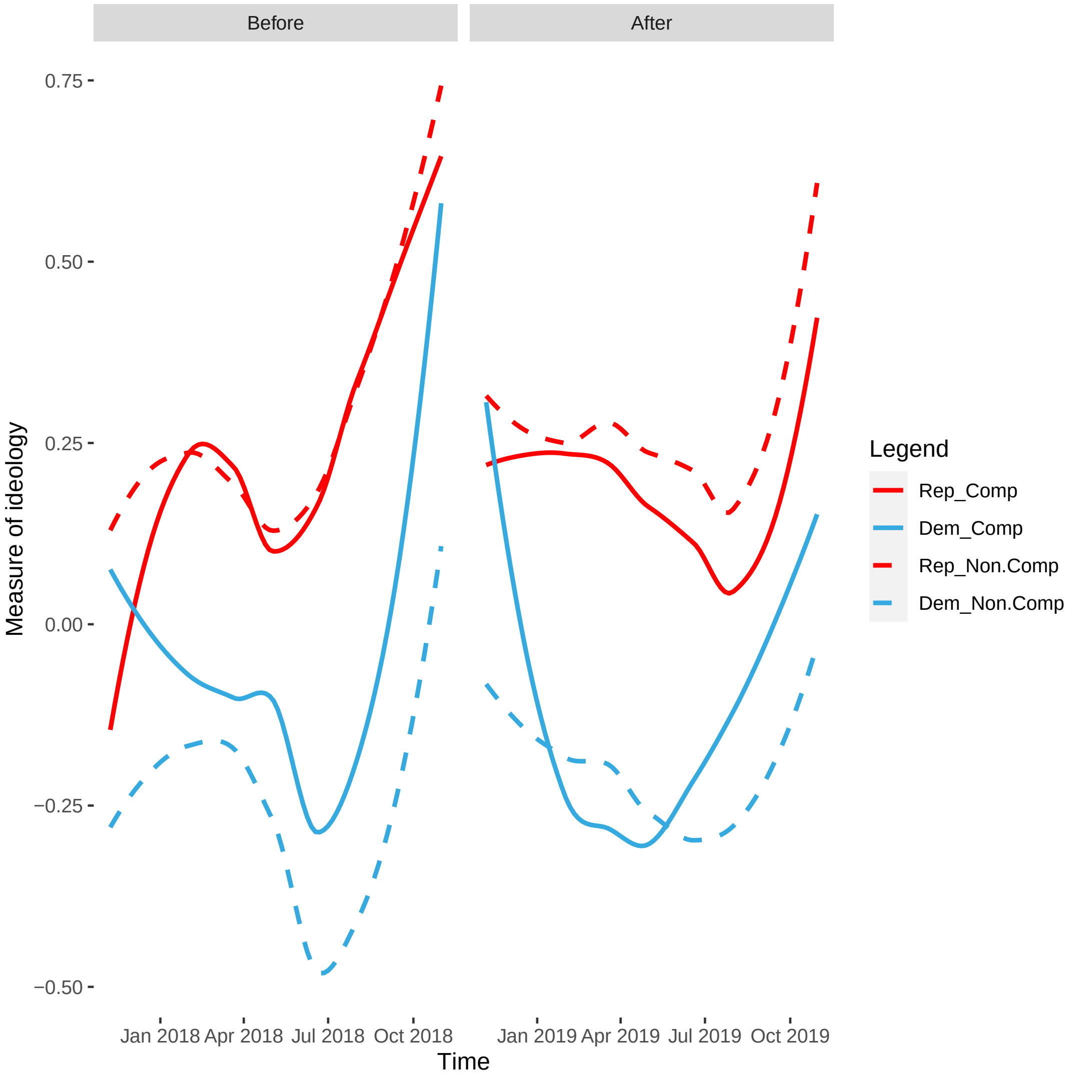
Notes: The top 5 domains that Democrats share more relative to Republicans as URLs. The number on the Y-axis shows the number of weeks that a particular domain has emerged the top domain.

Figure 1.14: Domains which have tweeted relatively more by Democrats comapred to Republicans



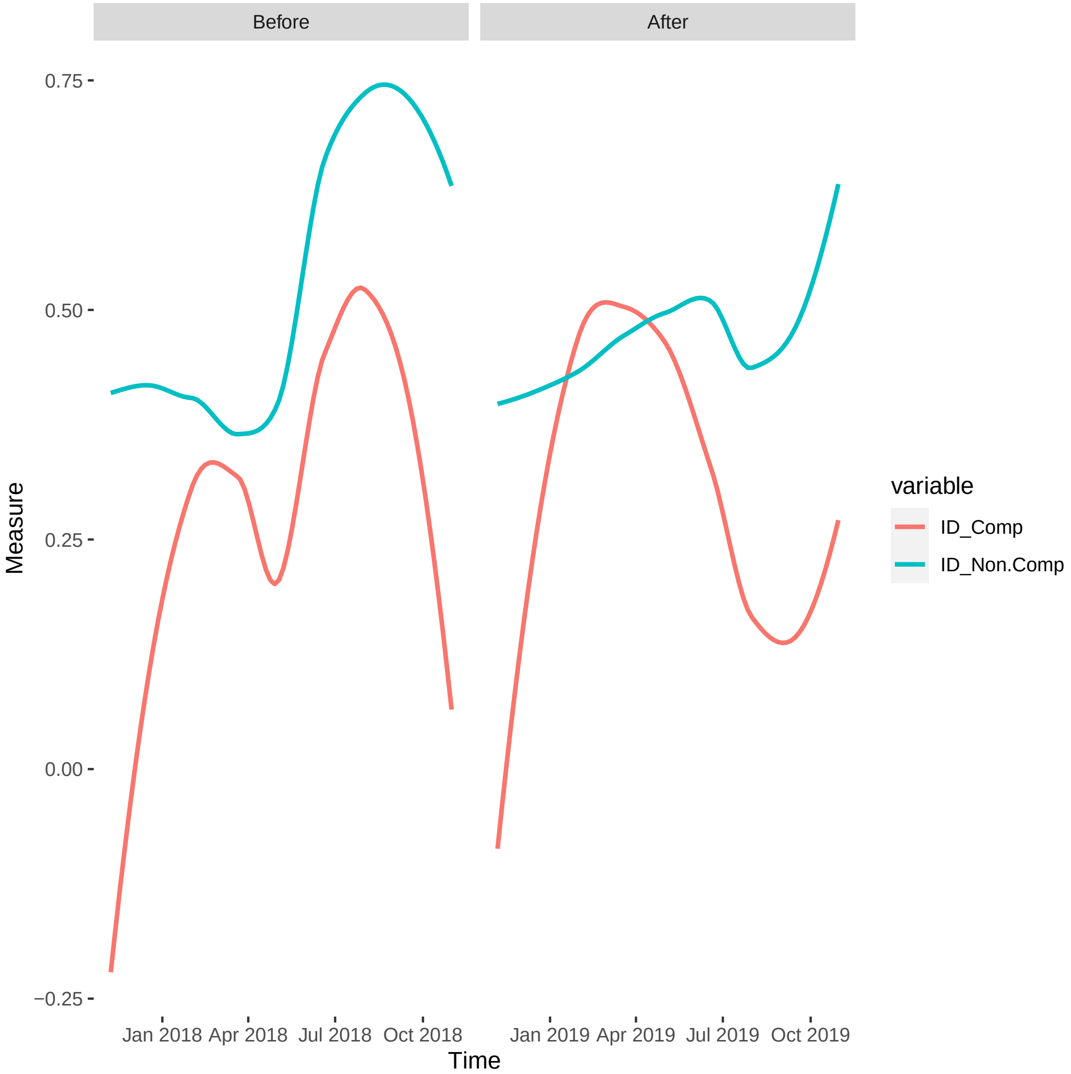
Notes: The top 5 domains that Democrats share more relative to Republicans as URLs. The number on the Y-axis shows the number of weeks that a particular domain has emerged the top domain.

Figure 1.15: Mean ideological score over time computed using Bayesian Ideal Point Estimation



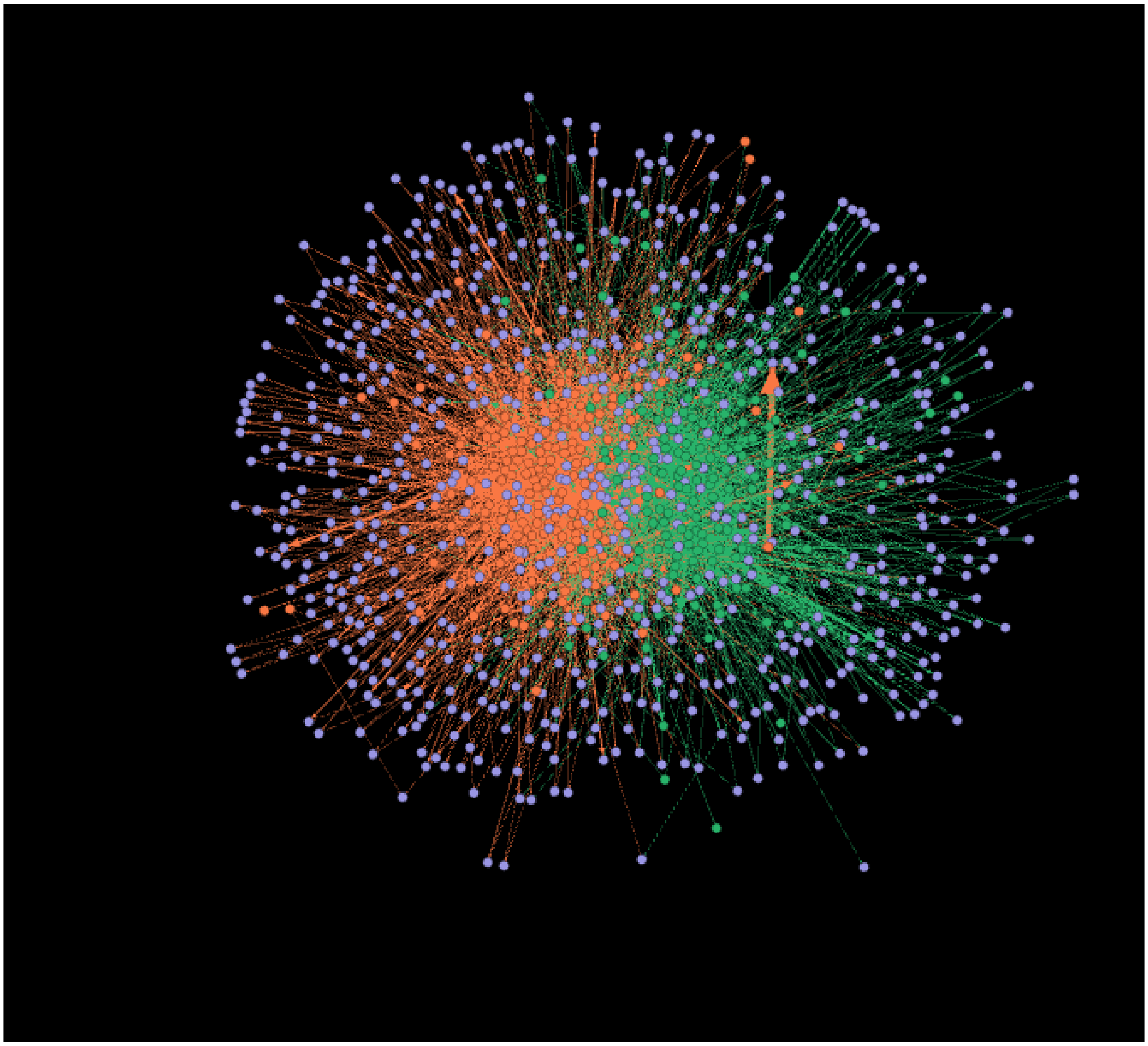
Notes: Rep_Comp and Dem_Comp shows the movement of ideological scores for Republicans and Democrats in competitive districts calculated using Bayesian Ideal Point estimation. Rep_Non-Comp and Dem_Non-Comp similarly shows the movement of ideological scores in non-competitive districts. The Before facet shows what the trend before the Nov 6^{th} midterm election, whereas the After facet shows the trend after the Nov 6^{th} midterm.

Figure 1.16: Mean ideological polarization over time computed using Bayesian Ideal Point Estimation.



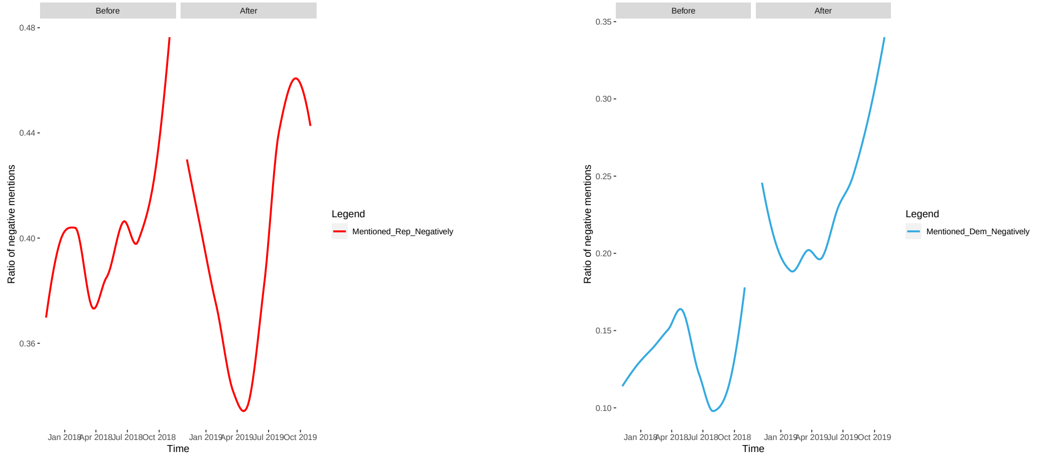
Notes: ID_Comp is the difference between the mean ideological scores of Republicans and Democrats in competitive districts whereas ID_Non-Comp is the difference in non-competitive districts. The Before facet shows what the trend before the Nov 6^{th} midterm election, whereas the After facet shows the trend after the Nov 6^{th} midterm.

Figure 1.17: Polarization in the mention network



Notes: The green colored dots represents Republicans, orange represents Democrats, and mauve represents non politicians

Figure 1.18: Trend of negative mentions by parties over time

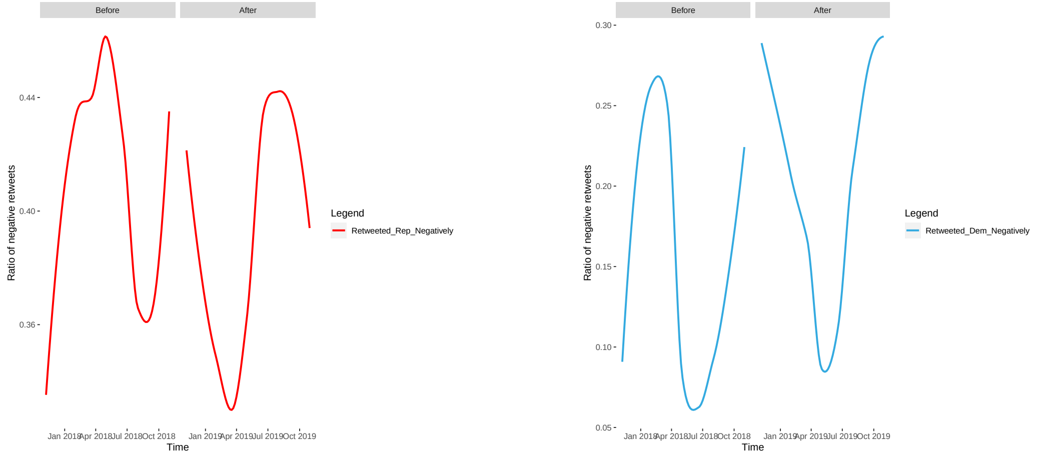


(a) Trend of negative mentions of Republicans by Democrats

(b) Trend of negative mentions of Democrats by Republicans

Notes: The Before facet shows what the trend before the Nov 6^{th} midterm election, whereas the After facet shows the trend after the Nov 6^{th} midterm.

Figure 1.19: Trend of negative retweets by parties over time



(a) Trend of negative retweets of Republicans by Democrats

(b) Trend of negative retweets of Democrats by Republicans

Notes: The Before facet shows what the trend before the Nov 6^{th} midterm election, whereas the After facet shows the trend after the Nov 6^{th} midterm.

Table 1.1: Top 20 hashtags used by the Democrats and Republicans

Panel A: One week before the election

Democrats	Republicans
#forthepeople, #vote	#betteroffnow, #jobsreport
#getcovered, *#electionday*	#taxreform, #taxcutsandjobsact
#shirleychisholm, #protectourcare	#halloween, #hurricanemichael
#investigatezinke, #unboughtunbossed	#ms01, #maga
#govote, #latinaequalpayday	#ar3, #happyhalloween
#cultureofcorruption, #latinaequalpay	#veterans, #az05
#showupforshabbat, #midterms2018	#jobs, #ga10
#marianastrong, #openenrollment	#al03, #mobileoffice
#aca, #yutu	#nc06, *#electionday*
#pda40, #goptaxscam	#mi06, #az08

Panel B: One week after the election

Democrats	Republicans
#veteransday, #getcovered	*#veteransday*, #campfire
#forthepeople, #woolseyfire	*#veterans*, #marinecorpsbirthday
#veterans, #protectmueller	#semperfi, *#veteransday2018*
#veteransday2018, #trump	#ar3, #findyourpark
#enoughisenough, #thousandoaks	#al03, #az05
#mueller, #daca	#ruralbizsummit, #betteroffnow
#counteveryvote, #hillfire	#floridarecount2018, #thankaveteran
#thxbirthcontrol, *#wwi*	#ms01, #ar4
#followthefacts, #yutu	#semperfidelis, *#wwi*
#gunviolence, #diwali	#nationaladoptionmonth, #nc06

Notes: **Panel A** shows the top 20 hashtags used by Republicans and Democrats one week before the election. The common hashtags are italicized and underlined. #electionday is the only common hashtag used by both Democrats and Republicans one week before the election. **Panel B** shows the top 20 hashtags used by Republicans and Democrats one week after the election. The common hashtags are #veteransday, #veterans, #veteransday2018 and #wwi.

Table 1.2: Regression Discontinuity Estimates for Hashtag Analysis

Panel A: RDD Estimates for Hashtag Similarity

	(1)	(2)	(3)	(4)	(5)
Degree	Hashtag$_{10}$	Hashtag$_{20}$	Hashtag$_{40}$	Hashtag$_{50}$	Hashtag$_{100}$
1	1.454	3.049***	3.389**	5.192***	12.048***
	(1.094)	(0.983)	(1.711)	(1.648)	(2.859)
2	3.586*	4.481***	6.278**	8.703***	17.272***
	(1.650)	(1.438)	(2.687)	(2.567)	(3.323)
3	3.296	4.330***	6.753**	9.189***	17.604***
	(1.723)	(1.617)	(3.190)	(2.963)	(4.249)
4	4.128*	4.163**	5.252	3.895	19.551***
	(1.851)	(1.774)	(3.838)	(3.820)	(4.862)
N	104	104	104	104	104

Panel B: RDD Estimates for $Inverse\_Score\_std$ estimates

	(1)	(2)	(3)	(4)	(5)
Degree	Inv_Score_std$_{10}$	Inv_Score_std$_{20}$	Inv_Score_std$_{40}$	Inv_Score_std$_{50}$	Inv_Score_std$_{100}$
1	3.050	6.967	0.074	-0.644	-3.344
	(12.147)	(8.164)	(5.578)	(6.207)	(3.808)
2	0.687	17.341	6.839	5.391	5.076
	(13.902)	(16.117)	(8.481)	(9.381)	(3.732)
3	10.732	19.138	3.968	6.733	11.191
	(24.855)	(18.158)	(9.968)	(10.767)	(9.210)
4	15.846	33.235	10.372	9.757	12.517
	(31.479)	(23.568)	(9.840)	(11.194)	(10.283)
N	104	104	104	104	104

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2] Hashtag similarity is defined as the number of common hashtags used by Democrats and Republicans conditional on the top hashtags used by them. $Inverse\_Score\_std$ is the inverse of distance between fraction of negative tweets for the Republicans and Democrats out of all the tweets that use a similar hashtag standardized for the number of topics. I perform a non-parametric RDD in Time using a data driven bandwidth selection method. The estimates are reported for 1, 2, 3 and 4 degrees of polynomials.

Table 1.3: Regression Discontinuity Estimates for Euclidean distance measures between topic distributions and word distributions

	(1)	(2)	(3)	(4)	(5)
Degree	Hellinger	KLD$_{DR}$	KLD$_{RD}$	JD$_{DR}$	JD$_{DLRL}$
1	-0.059	-0.060	-0.070	-0.018	-0.015*
	(0.051)	(0.053)	(0.062)	(0.015)	(0.008)
2	-0.070	-0.072	-0.085	-0.006	-0.012
	(0.062)	(0.067)	(0.077)	(0.028)	(0.011)
3	-0.095	-0.100	-0.108	-0.005	-0.004
	(0.065)	(0.068)	(0.080)	(0.035)	(0.018)
4	-0.081	-0.099	-0.105	-0.010	-0.007
	(0.086)	(0.085)	(0.118)	(0.041)	(0.018)
N	104	104	104	104	104

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2] *Hellinger* Distance and the KLD_{DR} and KLD_{RD} measure the distance between the topic distributions after implementing the LDA model on the content of the tweets. The JD_{DR} and JD_{DLRL} measures the distance between the word distributions used by the Democrats and Republicans in their tweets. I perform a non-parametric RDD in Time using a data driven bandwidth selection method. The estimates are reported for 1, 2, 3 and 4 degrees of polynomials.

Table 1.4: Regression Discontinuity Estimates for Euclidean Distance and Euclidean distance interacted with sentiment for dominant topic

Panel A: RDD Estimates for Euclidean Distance

	(1)
Degree	Score
1	-0.011
	(0.073)
2	-0.009
	(0.100)
3	-0.010
	(0.124)
4	0.042
	(0.170)
N	104

Panel B: RDD Estimates for sentiment augmented Euclidean distance

	(1)
Degree	Sum_Frac_Dis
1	0.023
	(0.212)
2	-0.205
	(0.402)
3	-0.317
	(0.620)
4	-0.269
	(0.604)
N	104

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2] *Score* is defined as the euclidean distance between the vectors of fraction of each topic(dominant topic in each tweet) for the Democrats and Republican tweets for each week.*Sum_Frac_Dis* is the euclidean distance for each topic between Republicans and Democrats, after controlling for the sentiment of the topics. I perform a non-parametric RDD in Time using a data driven bandwidth selection method. The estimates are reported for 1, 2, 3 and 4 degrees of polynomials.

Table 1.5: Regression Discontinuity Estimates for Bayesian Ideal Point(BIP) Estimates for politicians and Ideological Difference in competitive and non-competitive districts

Panel A: RDD Estimates for BIP Estimates for politicians

	(1)	(2)	(3)	(4)
Degree	Rep_Comp	Dem_Comp	Rep_Non-Comp	Dem_Non-Comp
1	-0.308	-0.264	-0.428	-0.122
	(0.399)	(0.450)	(0.305)	(0.346)
2	-0.259	0.268	-0.172	0.232
	(0.474)	(0.920)	(0.524)	(0.709)
3	0.102	0.297	0.236	0.245
	(0.746)	(1.032)	(0.735)	(0.684)
4	0.104	-0.109	0.376	0.142
	(0.841)	(1.683)	(0.853)	(1.100)
N	104	104	104	104

Panel B: RDD Estimates for Ideological Difference

	(1)	(2)
Degree	IdeologicalDiff_Comp	IdeologicalDiff_Non-Comp
1	-0.281	-0.278
	(0.295)	(0.189)
2	-0.343	-0.107
	0.471	(0.227)
3	-0.336	-0.008
	(0.580)	(0.218)
4	0.772	0.148
	(0.820)	(0.296)
N	104	104

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.
[2] Footnote 2: In **Panel A** *Rep_Comp* and *Dem_Comp* shows the mean ideological score for all Republicans and Democrats respectively in competitive districts.$Rep\_Non-Comp$ and $Dem\_Non-Comp$ shows the mean ideological score for all Republicans and Democrats respectively in non-competitive districts. In **Panel B** $IdeologicalDiff\_Comp$ and $IDeologicalDiff\_Non-Comp$ shows the difference between mean Republican and Democrat ideological scores in competitive and non-competitive districts respectively. I perform a non-parametric RDD in Time using a data driven bandwidth selection method. The estimates are reported for 1, 2, 3 and 4 degrees of polynomials.

Table 1.6: Regression Discontinuity Estimates for Mention Network Analysis

Panel A: RDD Estimates for Mentions by Democrats

	(1)
Degree	Rep_Negatively
1	-0.043
	(0.038)
2	-0.056
	(0.043)
3	-0.041
	(0.050)
4	0.044
	(0.068)
N	104

Panel B: RDD Estimates of mentions by Republicans

	(1)
Degree	Dem_Negatively
1	0.004
	(0.084)
2	0.041
	(0.042)
3	0.048***
	(0.134)
4	0.052***
	(0.116)
N	104

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2] In **Panel A** *Rep_Negatively* are shares of how many times Republicans are mentioned negatively relative to all mentions of Republicans by the Democrat politicians. In **Panel B** *Dem_Negatively* are shares of how many times Democrats are mentioned negatively relative to all mentions of Democrats by the Republican politicians. I perform a non-parametric RDD in Time using a data driven bandwidth selection method. The estimates are reported for 1, 2, 3 and 4 degrees of polynomials.

Table 1.7: Regression Discontinuity Estimates for Retweet Network Analysis

Panel A: RDD Estimates for Retweets by Democrats

	(1)
Degree	Rep_Negatively
1	-0.063
	(0.036)
2	-0.094**
	(0.048)
3	-0.108**
	(0.054)
4	-0.122**
	(0.056)
N	104

Panel B: RDD Estimates for Retweets by Republicans

	(1)
Degree	Dem_Negatively
1	0.019
	(0.029)
2	0.013
	(0.038)
3	-0.019
	(0.057)
4	-0.111
	(0.090)
N	104

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.
[2] In **Panel A** *Rep_Negatively* are shares of how many times Republicans are retweeted negatively relative to all retweets of Republicans by the Democrat politicians. In **Panel B** *Dem_Negatively* are shares of how many times Democrats are retweeted negatively relative to all retweets of Democrats by the Republican politicians. I perform a non-parametric RDD in Time using a data driven bandwidth selection method. The estimates are reported for 1, 2, 3 and 4 degrees of polynomials.

Table 1.8: OLS Sub-sample Estimates for Hashtag Analysis

Panel A: Estimates for Hashtag Similarity

	(1)	(2)
Metric	Before	After
Hashtag_{10}	-0.179	-0.179
	(0.1342)	(0.2712)
Hashtag_{20}	-0.333	0.0595
	(0.291)	(0.2576)
Hashtag_{40}	-0.536	-0.083
	(0.3323)	(0.381)
Hashtag_{50}	-0.8333**	0.0119
	(0.4123)	(0.4623)
Hashtag_{100}	-1.464**	-0.869
	(0.701)	(0.6209)
N	8	8

Panel B: Estimates of for$Inverse_Score_std$ estimates

	(1)	(2)
Metric	Before	After
$\text{Inv\_Score\_10}_{std}$	-2.279	-6.285
	(1.543)	(4.426)
$\text{Inv\_Score\_20}_{std}$	-0.514	-2.979
	(0.8651)	(1.087)
$\text{Inv\_Score\_40}_{std}$	-0.368	-0.348
	(0.6268)	(0.3469)
$\text{Inv\_Score\_50}_{std}$	-0.262	-0.587
	(0.7059)	(0.3062)
$\text{Inv\_Score\_100}_{std}$	-0.288	-0.486
	(0.1866)	(0.1836)
N	8	8

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2] Hashtag similarity is defined as the number of common hashtags used by Democrats and Republicans conditional on the top hashtags used by them. $Inverse\_Score\_std$ is the inverse of distance between fraction of negative tweets for the Republicans and Democrats out of all the tweets that use a similar hashtag standardized for the number of topics. The Before column shows the estimates for 8 weeks before the election whereas the After column shows the estimates for 8 weeks after the election.

Table 1.9: OLS Sub-sample estimates for Euclidean distances for topic distributions and word distributions

	(1)	(2)
Metric	Before	After
Hellinger	0.0055	-0.008
	(0.0052)	(0.0062)
KLD_DR	0.0064	-0.004
	(0.0057)	(0.0035)
KLD_RD	0.008	-0.004
	(0.0064)	(0.0035)
JD_DR	-0.006	0.0028
	(0.0017)	(0.0031)
JD_DLRL	-0.003	0.0008
	(0.0013)	(0.0028)
N	8	8

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2] *Hellinger* Distance and the KLD_{DR} and KLD_{RD} measure the distance between the topic distributions after implementing the LDA model on the content of the tweets. The JD_{DR} and JD_{DLRL} measures the distance between the word distributions used by the Democrats and Republicans in their tweets. The Before column shows the estimates for 8 weeks before the election whereas the After column shows the estimates for 8 weeks after the election.

Table 1.10: OLS Sub-sample Estimates for Euclidean Distance and Euclidean distance interacted with sentiment for dominant topic

Panel A: Estimates for Euclidean Distance

	(1)	(2)
Metric	Before	After
Score	0.0046	-0.025
	(0.0101)	(0.0128)
N	8	8

Panel B: Estimates for sentiment augmented Euclidean distance

	(1)	(2)
Metric	Before	After
Sum_Frac_Dis	0.0111	0.0262
	(0.0557)	(0.0259)
N	8	8

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2]Score is defined as the euclidean distance between the vectors of fraction of each topic(dominant topic in each tweet) for the Democrats and Republican tweets for each week. Sum_Frac_Dis is the euclidean distance for each topic between Republicans and Democrats after controlling for the sentiment of the topics. The Before column shows the estimates for 8 weeks before the election whereas the After column shows the estimates for 8 weeks after the election.

Table 1.11: OLS Sub-sample Estimates for Bayesian Ideal Point(BIP) Estimates for politicians and Ideological Difference in competitive and non-competitive districts

Panel A: Estimates for BIP Estimates for politicians

	(1)	(2)
Metric	Before	After
Rep_Comp	-0.055	-0.021
	(0.0756)	(0.0732)
Dem_Comp	-0.008	-0.107
	(0.1056)	(0.076)
Rep_Non-Comp	-0.04	-0.065
	(0.0664)	(0.0813)
Dem_Non-Com	-0.039	-0.04
	(0.0811)	(0.0626)
N	8	8

Panel B: Estimates for Ideological Difference

	(1)	(2)
Metric	Before	After
ID_Comp	-0.047	0.0865
	(0.0363)	(0.0511)
ID_Non-Comp	-0.001	-0.025
	(0.0375)	(0.0313)
N	8	8

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2] In **Panel A** *Rep_Comp* and *Dem_Comp* shows the mean ideological score for all Republicans and Democrats respectively in competitive districts.$Rep\_Non-Comp$ and $Dem\_Non-Comp$ shows the mean ideological score for all Republicans and Democrats respectively in non-competitive districts. In **Panel B** ID_Comp and $ID\_Non-Comp$ shows the difference between mean Republican and Democrat ideological scores in competitive and non-competitive districts respectively. The Before column shows the estimates for 8 weeks before the election whereas the After column shows the estimates for 8 weeks after the election.

Table 1.12: OLS Sub-sample Estimates for Mention Network Analysis

Panel A: Estimates for Mentions by Democrats		
	(1)	(2)
Metric	Before	After
Rep_Negatively	0.0134	-5E-04
	(0.0127)	(0.0134)
N	8	8

Panel B: Estimates of mentions by Republicans		
	(1)	(2)
Metric	Before	After
Dem_Negatively	0.0087	-0.036
	(0.0069)	(0.0198)
N	8	8

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.
[2] In **Panel A** *Rep_Negatively* are shares of how many times Republicans are mentioned negatively relative to all mentions of Republicans by the Democrat politicians. In **Panel B** *Dem_Negatively* are shares of how many times Democrats are mentioned negatively relative to all mentions of Democrats by the Republican politicians. The Before column shows the estimates for 8 weeks before the election whereas the After column shows the estimates for 8 weeks after the election.

Table 1.13: OLS Sub-sample Estimates for Retweet Network Analysis

Panel A: Estimates for Retweets by Democrats		
	(1)	(2)
Metric	Before	After
Rep_Negatively	-0.014	0.0198
	(0.0524)	(0.0566)
N	8	8
Panel B: Estimates for Retweets by Republicans		
	(1)	(2)
Metric	Before	After
Dem_Negatively	0.006	-0.015
	(0.0431)	(0.0682)
N	8	8

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2] In **Panel A** *Rep_Negatively* are shares of how many times Republicans are retweeted negatively relative to all retweets of Republicans by the Democrat politicians. In **Panel B** *Dem_Negatively* are shares of how many times Democrats are retweeted negatively relative to all retweets of Democrats by the Republican politicians. The Before column shows the estimates for 8 weeks before the election whereas the After column shows the estimates for 8 weeks after the election.

Chapter 2

High Frequency Political Polarization in 2019 Lok Sabha Elections from India. Evidence from Twitter Data

2.1 Introduction

Political partisanship affects not only voting outcomes but also impacts political and economic outcomes. Two different theories of short-run political partisanship have been proposed in the literature. One of the theories that assume that political parties are only motivated to hold office and win elections predict policy convergence to the median voter's preferences. However, other studies which assume ideologically motivated politicians or non-binding commitment devices predict policy divergence. Although some studies have tried to test these two hypotheses empirically, there has been no work done in the Indian context.

The political climate in India is going through some turbulent times. Popular news media and recent events suggest that there has been growing polarization in India in recent times. International media has also been cognizant of this development. An article by Masih and Staler, 2019 in The Washington Post talks about U.S style polarization spreading to India. They talk about how every issue is now viewed using a partisan lens, how social media has become a place for acrimony and how political debates have strained personal relationships. Gettleman et al. (2019) wrote an article in The New York Times about the rise of communal tensions, hate speech and trolling in India. These events have also contributed to increase in perceived polarization. In recent times, the Citizenship Amendment Act (CAA) passed on 11th December 2019, suspension of Article 370 in Jammu and Kashmir, viral videos of mob lynching events where Hindu extremists allegedly killed Muslims are some incidents that have caused widespread protests in the country. The passing of the CAA has seen some of the most consistent and large scale protests organized by civil societies that the country has witnessed in a long time. There have also been isolated incidents of violence which took a monstrous form on 23rd February 2020 in Delhi. It was one of the most devastating communal riots in decades. Fifty-three people are supposed to have died according to an NPR article published on 7th March 2020.

The spurt of such news in prestigious media organizations warrants an immediate study of the subject. Therefore, my research question is to look at political polarization between the politicians of the two main political parties in India in the event of the May 2019 election. My contribution in this paper is two-fold. First, I help test the policy convergence versus policy divergence hypothesis, using Indian data, which has not been attempted yet. This is the only paper to compute high-frequency ideological estimates of polarization in a time series fashion for India. Second, this work is crucial for the Indian context. Unlike some developed countries like the USA, which has the DW-Nominate scores to measure political ideology, India has no such measure, and this study could help fill that void.

Twitter data is an excellent source of data to identify this kind of variation for several reasons. Twitter data helps us gain insight into the political discourse at very short intervals, almost week by week. The previous chapter provides a detailed exposition on this.

To quantify the degree of polarization in the political system in India during the 2019 Lok Sabha elections, I look at the Twitter feed of politicians from two major national parties in India viz. the Bharatiya Janata Party henceforth referred to as BJP and the Indian National Congress henceforth referred to as the INC. India has a multi-party system along with a federal structure. Subsequently, it has many political parties both at the national and the regional level. However, it is difficult to do a comprehensive study of all such parties in one single paper. I choose to focus on the BJP and INC, the two major political parties with a pan India presence, as explained later in the paper. Another reason for choosing these two parties is that most small and regional parties generally enter into an alliance or a coalition with these two major parties to form the government or the opposition. Therefore, looking at the political polarization between these parties gives us a close albeit imperfect measure of political, ideological differences expressed through rhetoric.

To approximate the degree of polarization, I compute several metrics using hashtag analysis, topic modelling, Bayesian Ideal Point estimation, and mention network analysis of the BJP and INC politicians' Twitter feeds. I look at the Twitter feeds of politicians who won in the 2019 election and have a verified Twitter profile. Using a verified profile ensures that we do not find much cheap talk and add more legitimacy to the tweets.

2.2 Background And Literature Review

2.2.1 Institutional Background

India was a British colony until 1947. On 15th August 1947, India got its independence and was divided into India and Pakistan. Pakistan was later divided into Pakistan and Bangladesh in 1971.

Michelutti (2007) summarizes the history of Indian post-colonial politics in the following way. From Independence to 1967 the party system was dominated by the Congress party, also known as the Indian National Congress (INC). It was an inclusive secular party supported by upper castes, lower castes, and different religious and ethnic communities. From 1967 to 1993, although Congress remained dominant at the national level, they faced more meaningful opposition at the state and regional levels. The Janata Party, a coalition of opposition parties, took office in 1977, and then in 1989, the Congress was defeated again by a new coalition of the National Front/Janata Dal. Thirdly, in the late 1980s and early 1990s, the Indian party system moved from a one-dominant party system to a genuine competitive multi-party system. Regional parties began to capture a more significant share of votes, and they started to mobilize members of lower castes. Parallel to this trend (and sometimes in opposition), this period also saw the rise of Hindu nationalism and the Bharatiya Janata Party (BJP). BJP was in office at the national level from 1996 to 2004 either as a minority government or in coalition with regional parties. In 2004 the Bharatiya Janata Party lost the parliamentary elections, and Congress and its allies won. India experienced a very high sustained growth rate in the next decade, which was divided in two periods 2003-04 to 2007-08 and another brief period in 2009-10 and 2010-11, Dasgupta (2020). The economy however had a slowdown from 2011-12, Subramanian (2019). In 2014, the BJP won on anti-incumbency advantage. Scams caused by the Congress government, Kumar (2018) and the charisma of the current prime minister Narendra Modi, Jaffrelot (2015) made BJP win comfortably. In the current election of 2019, most people again rode high on the charisma on the Modi effect, BJP's organizational advantage, nationalist sentiment and expansive welfare policies, Chhibber and Verma (2019). An attack in Pulwama on Indian army by alleged terrorists led to the alleged killing of terrorists in Pakistan by the Indian Air Force. This also proved to be an essential factor for BJP's win in the 2019 election, where they won by a clear majority and did not need the support of any political party to form a coalition government.

It is clear from the discussion above that BJP and INC are the two major political parties in India who have a pan India presence. Table 2.1 shows the seat distribution by parties in different states of India from 1984 to 2019 at the national level. This covers a total of ten general elections or the Lok Sabha elections. The table shows that BJP and INC are the two parties who have won the highest number of seats for most states. Therefore, I look at political polarization between politicians of INC and BJP only, because these are the only two parties to have had a pan India presence.

India has a bicameral legislature system, with two Parliament houses, the Rajya Sabha (the Upper House) and the Lok Sabha (the Lower House). The Lok Sabha members are elected directly by the people, whereas the members of the Rajya Sabha are appointed. The Lok Sabha has more power than the Rajya Sabha. The maximum strength of the House envisaged by the Constitution is 552, which is made up by the election of up to 530 members to represent the States, up to 20 members to represent the Union Territories and not more than two members of the Anglo-Indian Community to be nominated by the Hon'ble President, if, in his/her opinion, that community is not adequately represented in the House. The total elective membership is distributed among the States so that the ratio between the number of seats allotted to each State and the population of the State is, so far as practicable, the same for all States.

The winning party chooses the Prime Minister. The Prime Minister chooses his/her council of ministers entrusted with ministries, such as Ministry of Human Resources and Development, Finance Ministry and others.

The 2019 Indian general election was held in seven phases from 11 April to 19 May 2019 to constitute the 17th Lok Sabha. Five hundred forty-three seats were contested. The votes were counted, and the result declared on 23 May 2019. About 910 million people were eligible to vote, and voter turnout was over 67 per cent – the highest ever, and the highest ever participation by women voters. The Bharatiya Janata Party won 303 seats, further increasing its substantial majority and the BJP-led National Democratic Alliance (NDA) won 353 seats. The BJP won 37.36 percent of votes. The Indian National Congress won 52 seats, and the INC-led United Progressive Alliance won 91. Other parties and their alliances won 98 seats.

2.3 Data

In this paper, I collect the official Twitter handles of all politicians from the BJP and INC who have won in the 2019 election. India does not have an official web page for elected politicians. Therefore, I searched for the politician's name and choose the official handle after matching the face, credentials and ensuring that it is a verified account. There are 120 official BJP handles, and 15 official INC handles in my data-set. I collect weekly tweets from 12th April 2018 to 20th May 2020. In essence, I collect data for one year before the starting date of the election and one year after the election's end date. The number of Twitter handles in my data-set for approximately 40 percent of the BJP politicians and 29 percent of the INC politicians who won the 2019 election.

Since India is a country of many languages, politicians use several different languages to communicate with the electorate. However, Hindi and English are the major languages of communication, especially in central elections. For BJP, whose main voter base is in northern India where the majority of Hindi speaking population resides, Hindi seems an obvious language choice. I collect tweets made in English as well as tweets made in Hindi between 12th April 2018 to 20th May 2020. Figure 2.1 shows the number of active politicians every week. An active politician is defined as someone who has tweeted in that week. As shown in the graph, there is an almost similar number of active BJP politicians tweeting in Hindi and English. In contrast, in INC, the politicians mostly tweeted in English with only a couple of politicians tweeting in Hindi.

Figure 2.2 shows the total tweets by BJP and INC politicians made every week. Again, as can be seen through the figure, BJP politicians make an almost equal number of tweets in Hindi and English compared to INC politicians who make most of their tweets in English. Figure 2.3 shows the number of tweets made on average by an active politician in a particular week. The INC politicians make almost the same number of average tweets in English as the BJP makes in English and Hindi but hardly tweet in Hindi.

2.4 Computation of Ideological Estimates

I perform the same analysis for the tweets that I performed for the US data. A detailed exposition of my motivation for these techniques can be found in the first chapter.

2.4.1 Hashtag Analysis

To start off with the computation of metrics of polarization, I look at the similarity in hashtags used. Hashtag similarity is defined as the number of common hashtags used by BJP and INC politicians conditional on the top hashtags used by them. To compute the hashtag similarity, I proceed in the following way: First, I extract the top 40 hashtags used by BJP in a week. Let us denote this set of hashtags by B_{40}. Second I extract the top 40 hashtags used by INC in a week.[1] Let us denote this set of hashtags by I_{40}. I then compute the number of similar hashtags between the sets B_{40} and I_{40}. Let us denote this by $Hashtag_{40}$. In other words,

$$Hashtag_{40} = n(B_{40} \cap I_{40}), \tag{2.1}$$

where $n(.)$ denotes the cardinal number. I also compute $Hashtag_{10}$, $Hashtag_{20}$, $Hashtag_{50}$, $Hashtag_{100}$ for robustness checks.

[1] I convert all the hashtags to lower case because sometimes the same hashtags can be written in different cases.

Figure 2.4 shows the trend of these metrics over the election cycle. The *Before* segment refers to the period before the election, the *Election* segment refers to the weeks during which election was being conducted and the *After* segment refers to the period after the election. I use the English tweets for this figure. The figure looks similar when I use the English and Hindi tweets as shown in the Appendix. Number of common hashtags increases as the election approaches, except for the top 100 hashtag. This gives some evidence that unlike the US case, politicians from BJP and INC talk about similar agendas as the election approaches. However, the pattern is less consistent than it was for the US election. This probably happens because Indian politicians are not very well versed in using hashtags, compared to their US counterparts. However, taking recent events in India into consideration they seem to be getting better and more tactical in the usage of hashtags with each passing day.

2.4.1.1 Sentiment augmented Euclidean Distance between hashtags

To combine the hashtag analysis with the sentiment analysis, I compute the distance between fraction of negative tweets for the BJP and INC politicians out of all the tweets that use a similar hashtag. Again, a detailed motivation of this technique is discussed in the first chapter.

To apply the technique to the US data, assume that there are s common topics in the top 40 hashtags used by BJP and INC. Therefore, the length of $Hashtag_{40}$ which we have already defined is s. I now construct two vectors I_{40_s} and B_{40_s}. Let the first element of I_{40_s} be denoted by $I_{40_s}(1)$. Then,

$$I_{40_s}(1) = \left[\frac{n\begin{pmatrix} Tweets\ by\ INC\ which\ contain\ the\ first\ hashtag \\ \quad in\ top\ 40\ hashtags\ and\ have\ a\ negative\ sentiment \end{pmatrix}}{n\begin{pmatrix} Tweets\ by\ INC\ which\ contain\ the\ first\ hashtag \\ \hfill in\ top\ 40\ hashtags \end{pmatrix}} \right], \quad (2.2)$$

where n denotes the cardinal number. I similarly compute all the elements for I_{40_s} and B_{40_s}, and find the Euclidean Distance between these two vectors. This is denoted by $Score_{40}$, where $Score_{40}$ is defined as follows:

$$Score_{40} = d(I_{40_s}, B_{40_s}) = \sqrt{\begin{matrix} (I_{40_s}(1) - B_{40_s}(1))^2 + (I_{40_s}(2) - B_{40_s}(2))^2 + ... \\ \hfill + (I_{40_s}(s) - B_{40_s}(s))^2 \end{matrix}}. \quad (2.3)$$

After this, I standardize the scores by the number of common hashtags by dividing $Score$ by the square root of the number of common hashtags. For example, $Score_{40_{std}}$ is computed as follows

$$Score_{40_{std}} = \frac{Score}{\sqrt{s}}. \quad (2.4)$$

I similarly also compute $Score_{10_{std}}$, $Score_{20_{std}}$, $Score_{50_{std}}$ and $Score_{100_{std}}$. The reason behind focusing only on negative sentiments is explained in the previous chapter.

Another point to note is that in doing the actual analysis I use the inverse of $Score_{10}$ and $Score_{10_{std}}$, which I refer to as Inv_Score_10 and $Inv\_Score\_10_{std}$. This is done because if there are no common hashtags for any of the groups, then distance would be calculated as 0, but that does not make sense because a distance close to 0 implies no polarization whereas 0 common hashtags does not imply the same. To resolve this ambiguity I take the inverse of the score, such that a high score means less polarization and low score means high polarization. When there are no common hashtags the metric is set to a value of 0, as no common hashtags imply the greatest degree of polarization.

Figure 2.5 shows the patterns in the inverse of the standardized scores. The graph shows that as we approach the election the inverse of the Euclidean distance increases, which means that conditional on using the same hashtags BJP and INC politicians use them with similar sentiments, although there is some discrepancy between the different lines. The sentiments used in the top hashtags become similar as the election approaches, although there is some inconsistency in the pattern. But broadly speaking, it seems that Indian politicians are not using extremely divisive hashtags in their top hashtags especially in the top 10, 20 and 40 top hashtags. This probably is an artifact of the fact that Indian politicians are not heavy hashtag users in the first place.

2.4.2 Topic Modelling

To perform topic modelling, I apply the model of Latent Dirichlet Allocation (LDA) to my corpora of tweets. The procedure of implementing a LDA is described in detail in the first chapter of the dissertation. Whereas I applied the LDA model to Democrat and Republican tweets in the first chapter, I apply the same technique to BJP and INC tweets here.

2.4.2.1 Computation of Distance Metrics

I compute measures of similarity and dissimilarity between the two topic probability distributions obtained after running the trained LDA model (trained on the pooled tweets of BJP and INC politicians) separately on BJP and INC tweets. The three measures which are used to measure the distance are the Hellinger distance, Kullback-Leibler divergence and the Jaccard distance which are explained in the previous chapter. The Hellinger distance and the Kullback-Leibler divergence is used to measure the distance between the topic probability distributions, whereas the Jaccard distance measures the distance between the word distributions.

Figure 2.6 shows the patterns in these distance based metrics over the election cycle. The Hellinger distace is denoted by Hellinger whereas the Kullback-Leibler Divergence is denoted by KLD_BI and KLD_IB because it is not a symmetric distance measure. The Jaccard Distance between the raw word distributions used by BJP and INC is denoted by J_BI whereas the distance between the vectors containing list of words used by BJP and INC is denoted by JL_BI. As can be seen in the figure the distance between the topic distributions as well as the word distributions increases as we approach the election. This suggests that politicians talk about different topics and use different words as we approach the election.

2.4.2.2 Euclidean Distance for the dominant topic

The LDA model assigns each tweet to multiple topics. For the next part of the analysis, I find the dominant topic for each tweet and assign the tweet to that particular topic. I calculate the Euclidean Distance between the vectors of fraction of tweets devoted to each topic by the BJP and INC for each week. For example, if there are s topics in a particular week, I obtain two vectors I_s and B_s for that week. Both these vectors have s elements. Let the first element of I_s vector be denoted by $I_s(1)$. Then,

$$I_s(1) = \frac{n(Tweets\,by\,INC\,which\,belong\,to\,topic\,1)}{n(Total\,tweets\,by\,INC)}, \tag{2.5}$$

where n denotes the cardinal number. I similarly compute all the elements of I_s and B_s and find the Euclidean Distance between these two vectors. This distance is denoted by $Score$ which is calculated as follows.

$$Score = d(I_s, B_s) = \sqrt{\begin{aligned}(I_s(1) - B_s(1))^2 + (I_s(2) - B_s(2))^2 + ...\\ + (I_s(s) - B_s(s))^2\end{aligned}} \tag{2.6}$$

I also compute a standardized score which is referred to as $Score_{std}$. This is computed as follows.

$$Score_{std} = \frac{Score}{\sqrt{s}} \tag{2.7}$$

$Score$ helps in understanding the between topic variability in the dominant topics used by BJP and INC. A low value of $Score$ implies that both the parties devote similar weights to the various dominant topics they use in their tweets whereas a higher value of $Score$ implies that they talk about different topics.

Figure 2.7 shows the patterns in the $Score$ and $Score_{std}$ variable. With the approach of the election, $Score_{std}$ increases, suggesting that politicians talk about different topics with the approach of the election, suggesting an increase in polarization.

2.4.2.3 Euclidean Distance Interacted with Sentiments

The Euclidean Distance helps us understand the pattern in the usage of topics by BJP and INC but provide no insight into how these topics are being used. To make more sense of the intent of the content used for each topic, I perform sentiment augmented content analysis and compute two types of metrics where sentiment is interacted with the topic. These measures give us a low value if BJP and INC politicians talk about the same topic with similar sentiments, and gives us a high value if they talk about the same topic using different sentiments. The two approaches that I take are as follows.

In the first case, I compute the fraction of positive, negative and neutral shares for each topic for BJP and INC separately, then compute the euclidean distance for each topic between BJP and INC politicians, and add the distances for all the topics. For example, if there are s topics in a particular week, I compute 6 vectors $Positive_{I_s}$, $Negative_{I_s}$, $Neutral_{I_s}$, $Positive_{B_s}$, $Negative_{B_s}$, $Neutral_{B_s}$ with s elements each. The first element for $Positive_{I_s}$, denoted by $Positive_{I_s}(1)$ is defined as follows,

$$Positive_{I_s}(1) = \frac{n\begin{pmatrix} Tweets\ by\ INC\ which\ belong\ to\ topic\ 1 \\ \hfill and\ have\ a\ positive\ sentiment \end{pmatrix}}{n(Total\ tweets\ by\ INC\ which\ belong\ to\ topic\ 1)}, \tag{2.8}$$

where n denotes the cardinal number. I similarly compute all the other vectors. My first measure Sum_Frac_Dis is defined as follows,

$$Sum\_Frac\_Dis = \frac{\begin{array}{r} d(Positive_{I_s}, Positive_{B_s}) + d(Negative_{I_s}, Negative_{B_S}) \\ + d(Neutral_{I_s}, Neutral_{B_s}) \end{array}}{}, \tag{2.9}$$

where $d(.,.)$ denotes the Euclidean distance between two vectors.

I also computed a standardized version of Sum_Frac_Dis which is defined as follows,

$$Sum\_Frac\_Dis_{std} = \frac{Sum\_Frac\_Dis}{\sqrt{n}} \tag{2.10}$$

Second, instead of computing the fractions of positive tweets for a particular topic for each group, I use the intensity of the sentiment to derive the metric for Euclidean distance. Therefore, instead of counting the number of positive, negative or neutral tweets for each group, I compute the intensity of positivity, negativity and neutrality in the tweets. For s topics, I again compute the six vectors $Positive_{I_s}$, $Negative_{I_s}$,

$Neutral_{I_s}$, $Positive_{B_s}$, $Negative_{B_s}$, $Neutral_{B_s}$. The first element for $Positive_{I_s}$, denoted by $Positive_{I_s}(1)$ is defined as follows,

$$Positive_{I_s}(1) = \begin{array}{c} Mean\ value\ of\ positive\ score\ for\ tweets\ by\ INC \\ which\ had\ a\ positive\ sentiment\ and\ belonged\ to\ topic\ 1 \end{array} \tag{2.11}$$

The metric, Sum_Dis is then calculated as follows,

$$Sum\_Dis = \begin{array}{r} d(Positive_{I_s}, Positive_{B_s}) + d(Negative_{I_s}, Negative_{B_s}) \\ + d(Neutral_{I_s}, Neutral_{B_s}) \end{array}, \tag{2.12}$$

where $d(.,.)$ denotes the Euclidean distance between two vectors.

I also computed a standardized version of Sum_Dis which is defined as follows,

$$Sum\_Dis_{std} = \frac{Sum\_Dis}{\sqrt{n}} \tag{2.13}$$

Figure 2.8 shows that sentiment augmented topic distance falls as we approach the election as shown by the patterns in Sum_Frac_Dis and $Sum\_Frac\_Dis_{std}$. Unlike the $Score$ and $Score_{std}$ metric, the value of Sum_Frac_Dis and $Sum\_Frac\_Dis_{std}$ falls as approach the election. This suggests that BJP and INC politicians use similar sentiment to talk about common topics as we approach the election. We get the same trends when we use the intensity of emotions. The figure is shown in the Appendix.

2.4.3 Bayesian Ideal Point Estimation

The third way by which I try to capture the short-run polarization is through computation of ideological estimates of the politicians using URLs that they share in their tweets by the method of Bayesian ideal point estimation. The method is explained in details in the previous chapter.

Figure 2.9 shows that as election approaches both INC and BJP become more like BJP in their URL sharing behavior. Although the distinction between liberal or conservative is not so clear in the Indian context, BJP is considered to be the more conservative party, and hence one can say that both parties become more conservative as election approaches. This is similar to the U.S context. However, after the election, ideological estimates drop and begin to increase again. This is dissimilar to the U.S context where the ideological estimates drop and keeps decreasing for sometime after the election. Figure 2.10 shows that ideological polarization increases as election approaches in India, and this in stark contrast to the U.S data where the polarization decreases as election approaches

2.4.4 Mentions Network Analysis

For the fourth part of ideological metric, I calculate the polarization in the mentions network in my data set. This gives us an idea about the affective polarization within the network. To compute the degree of affective polarization in this network, I compute the shares of how many times BJP politicians mention other INC politicians negatively over the span of 104 weeks, and vice-versa.

Figure 2.11a shows that negative mentions of BJP politicians by INC politicians decreases as the election approaches. Figure 2.11b shows that negative mentions of INC politicians by BJP politicians also decreases as the election approaches. This is also in contrast to the US data, where the negative mentions increases as the election approaches.

2.5 Empirical Analysis

2.5.1 Regression Discontinuity Analysis

I use the non-parametric Regression Discontinuity Design as explained in the first chapter to find if there is a significant discontinuity before and after the election. Since the elections in India are conducted for over a month, I look at what happens to the estimates before and after the election ignoring the election period. This could possibly have the effect of making significant effects undetectable but since the number of politicians with a verified Twitter profile in India is so less, I do not have the luxury to subset the data by election dates.

2.5.1.1 Results

For the RDD estimates, I find a significant decrease in common hashtags for the top 10 hashtags and and a significant increase in common hashtags for the top 50 hashtags as shown in Table 2.2. There is also a significant drop in similarity of words used before and after the election as shown in Table 2.4. Also, we find that the BJP politicians become significantly less conservative after the election and ideological difference between BJP and INC politicians falls significantly after the election as shown in Table 2.5 Panel A and Table 2.5 Panel B respectively. None of the other estimates have any significant discontinuities as shown in Tables 2.2 - 2.6.

2.5.2 OlS Estimates for sub sample

To look at what happens close to the election, I employ a similar strategy of focusing close to the election. Although for the US election, I chose 8 weeks to make sure that I did not capture any effect from the primaries. India does not have any election analogous to the primaries, and hence the choice of 8 weeks is not so much a strategic choice in India, as it is to just maintain comparability with the US. Therefore, to look at what happens before and after the election I zoom into 8 weeks before and after the election.

2.5.2.1 Results

Table 2.7 shows that slope estimates for similarity in hashtags increases as we approach the election, except for the top 100 hashtags, although the estimates are not significant. The slope estimates for the inverse score metrics however increase as we approach the election meaning that conditional on using the same hashtags BJP and INC politicians use similar sentiments to talk about those hashtags. Therefore both these estimates suggest a decrease in polarization as they approach the election.

The Euclidean Distance between the topic probability distributions and word distributions increases as the election approaches as shown in Table 2.8. Even on focusing on the most dominant topic for each tweet the Euclidean distance between the share of tweets devoted to each topic by BJP and INC politicians increases as the election approaches as shown in Table 2.9, suggesting that politicians do become more polarized with respect to their agenda setting behaviour with the approach of the election. However, when the topic analysis is complemented with sentiment analysis, the sentiment augmented Euclidean Distance decreases as we approach the election. This is similar to the results observed in the U.S context and shows that politicians diverge in their agenda setting behaviour but converge in the sentiments used for a particular topic.

The ideological scores for the both the BJP and INC politicians becomes significantly more conservative as the election approaches as shown in Table 2.10, whereas the ideological difference increases as the election approaches as shown in Table 2.10.

Negative mentions of BJP politicians by INC politicians and INC politicians by BJP politicians decreases as we approach the election.

The combined pieces of evidence suggests similar patterns in India as we found in USA. Politicians diverge in their agenda setting behaviour while converging in the agenda-item specific positioning. Some points of departure from the US results are that ideological difference as measured by Bayesian Ideal Point Estimation increases with the approach of the election in India while it decreases in USA. The hashtag analysis for India also has some contradictory results depending on whether one is looking at the top 10, 20, 40, 50 or 100 hashtags. The reason for this is the relatively lower use of hashtags in India by mainstream politicians leading to a lot of heterogeneity in the numbers. Another point of departure between the US and India is that conditional on using the same hashtags, politicians from opposing political parties in US use different sentiments to talk about the same hashtags whereas politicians from opposing political parties use similar sentiments to talk about the same hashtags with the approach of the election.

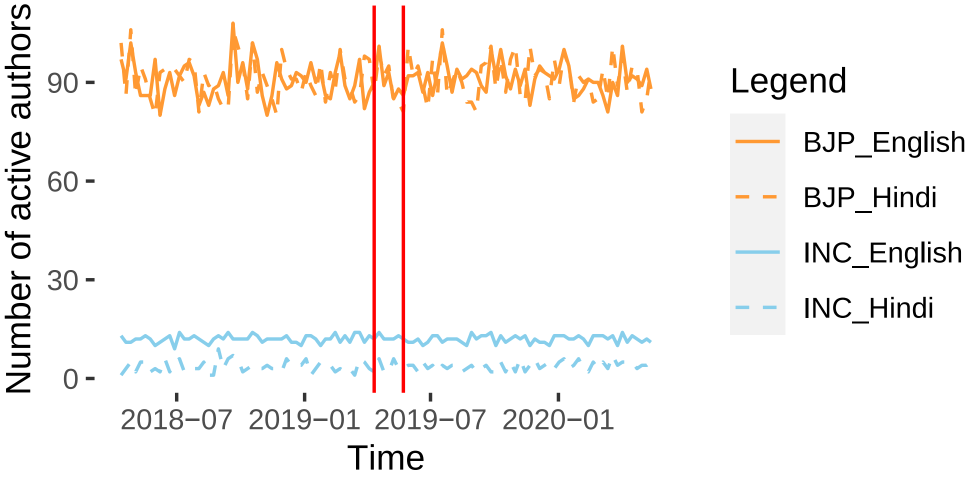
2.6 Conclusion

In this paper I collect tweets from BJP and INC politicians in India one year before and after the 2019 Lok Sabha elections. The tweets are then used to compute measures of ideological polarization.

My results show broadly that while politicians diverge in the agendas that they talk about with the approach of the election, they converge in the sentiments they use to talk about those topics. The results of the analysis in the Indian context are consistent with the results from the US context while looking at the content of the tweets. However, the results from the hashtag analysis, cited media ideology analysis as well as the mention network analysis is different from the US context.

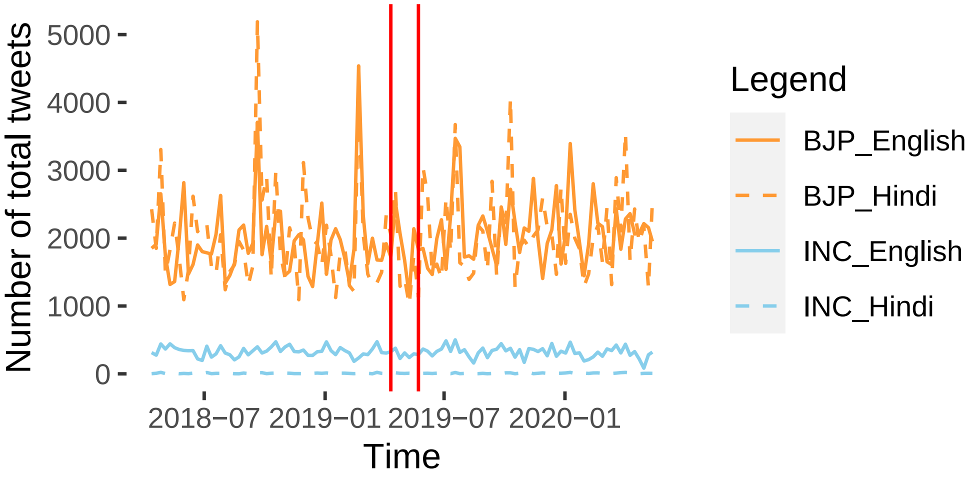
2.7 Figures and Tables

Figure 2.1: Total number of active politicians in each week.



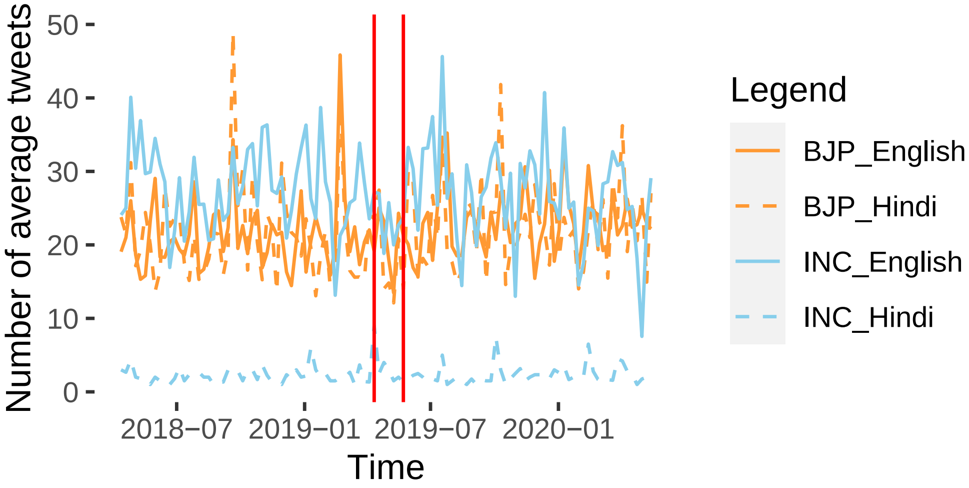
Notes: The number of active authors (authors who have tweeted atleast one tweet in a week) are counted for each week. The red vertical lines indicate the election.

Figure 2.2: Total number of tweets in each week.



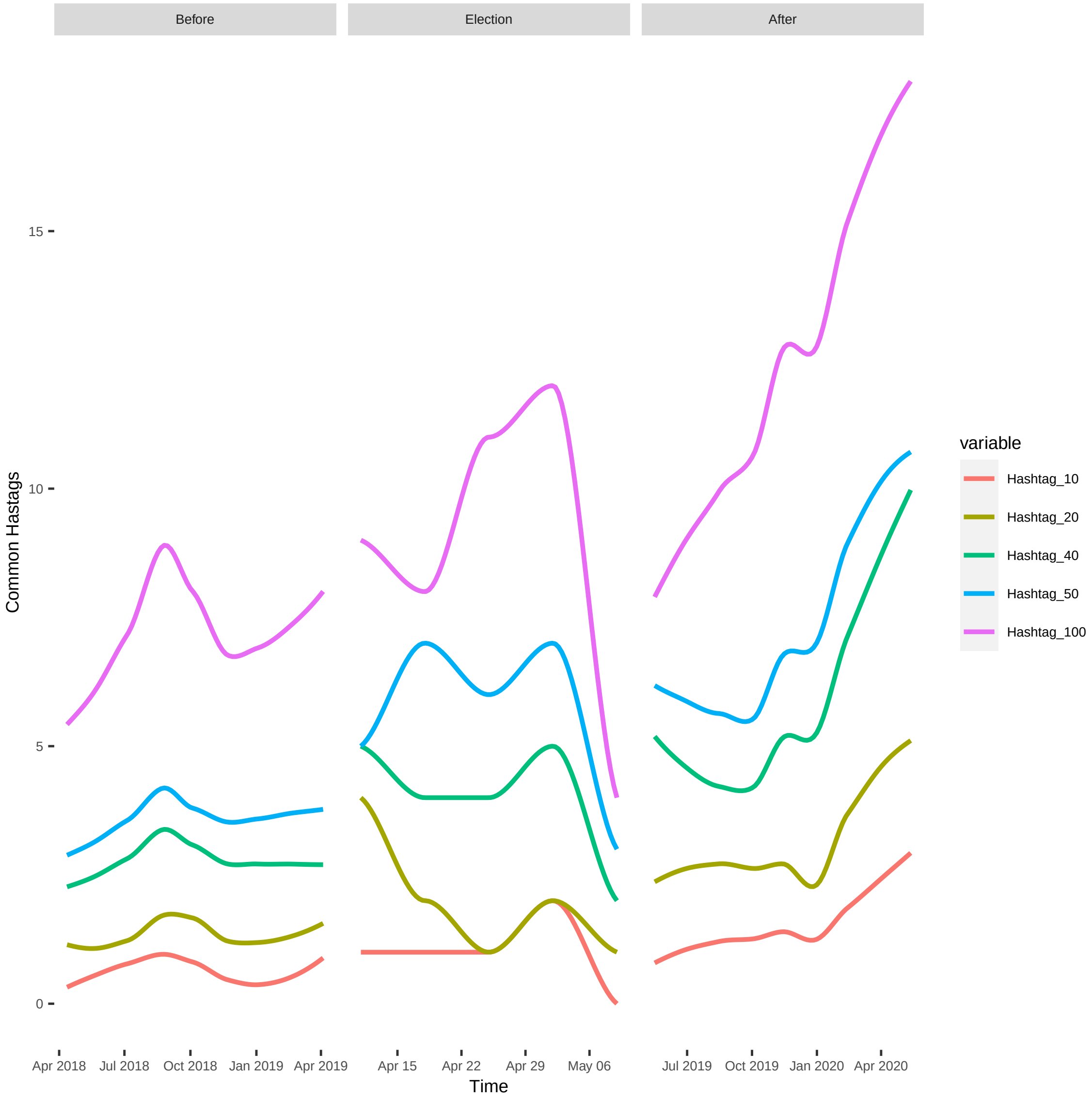
Notes: The total number of tweets are plotted for each week. The red vertical lines indicate the election.

Figure 2.3: Number of average tweets in each week.



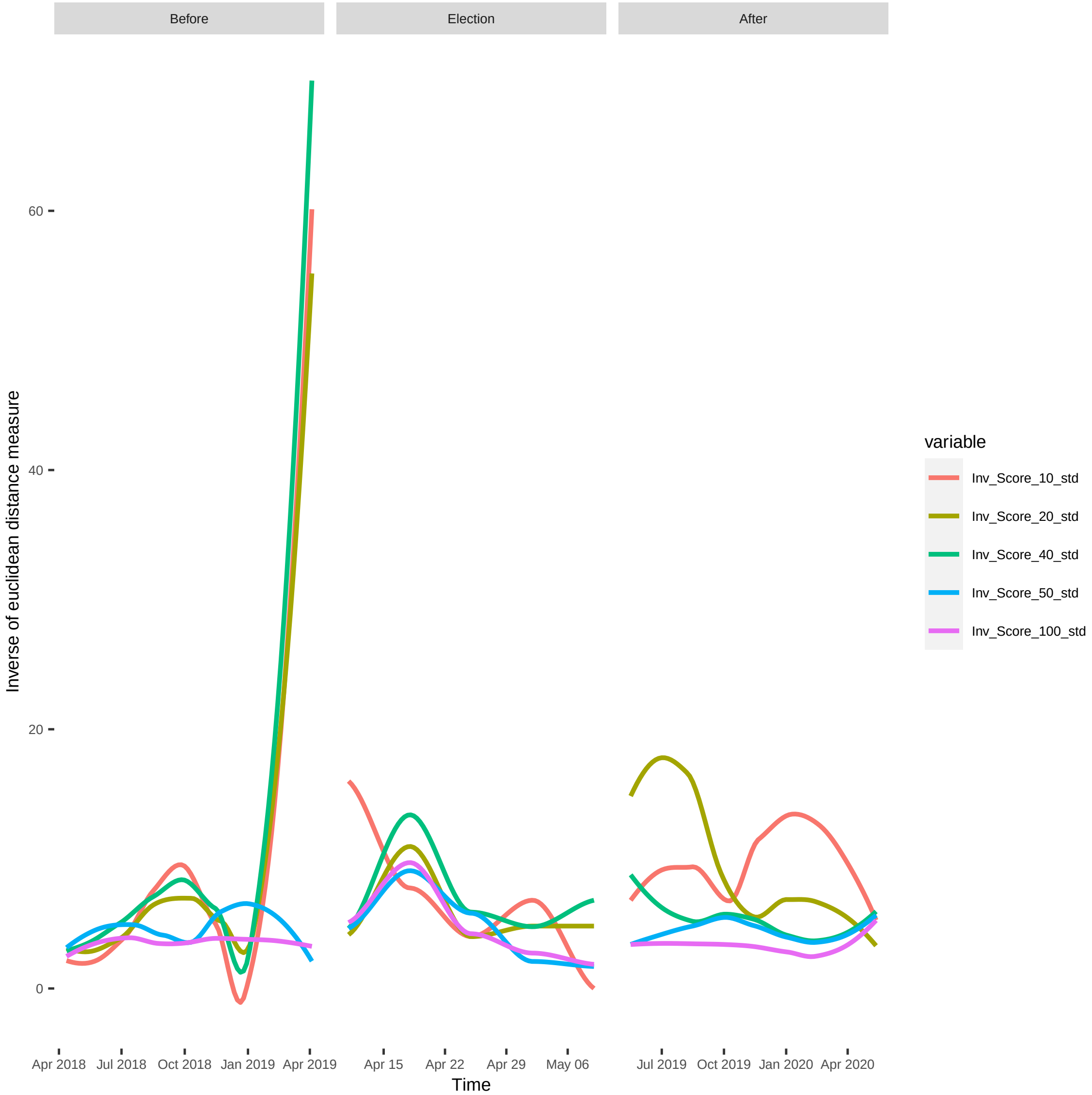
Notes: The number of tweets sent out on average each week. The numbers plotted in this graph is obtained by dividing the total number of tweets by the number of active authors. The red vertical line indicates the election.

Figure 2.4: Similarity in hashtags over time



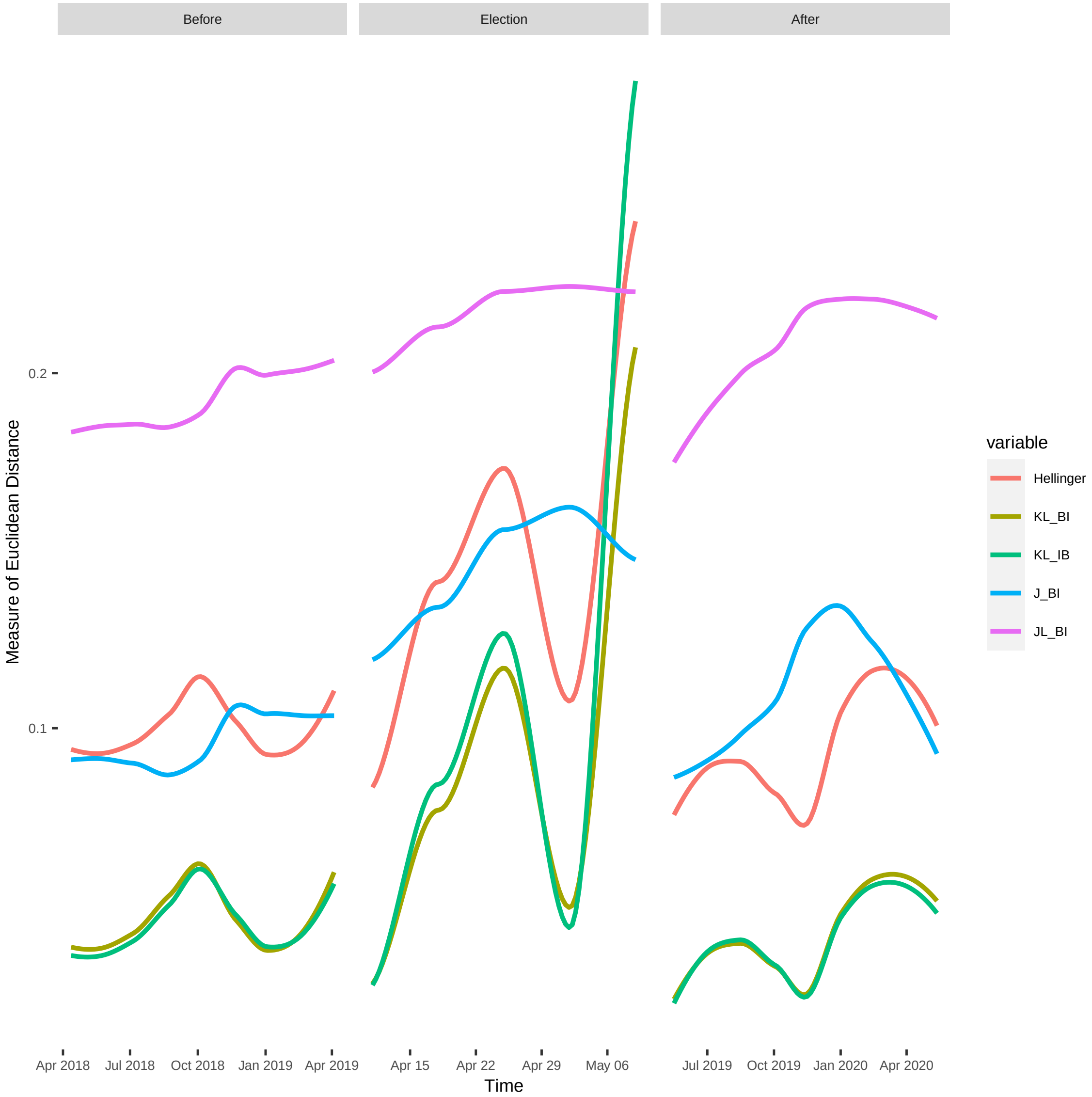
Notes: The figure plots the number of common hashtags used in the top 10, 20, 40, 50 and 100 hashtags used by BJP and INC politicians in English tweets. The Before facet shows the pattern before the election, the Election period shows the pattern during the election whereas the After facet shows the pattern after election.

Figure 2.5: Trend of inverse of standardized Euclidean Distance of hashtags interacted with sentiments



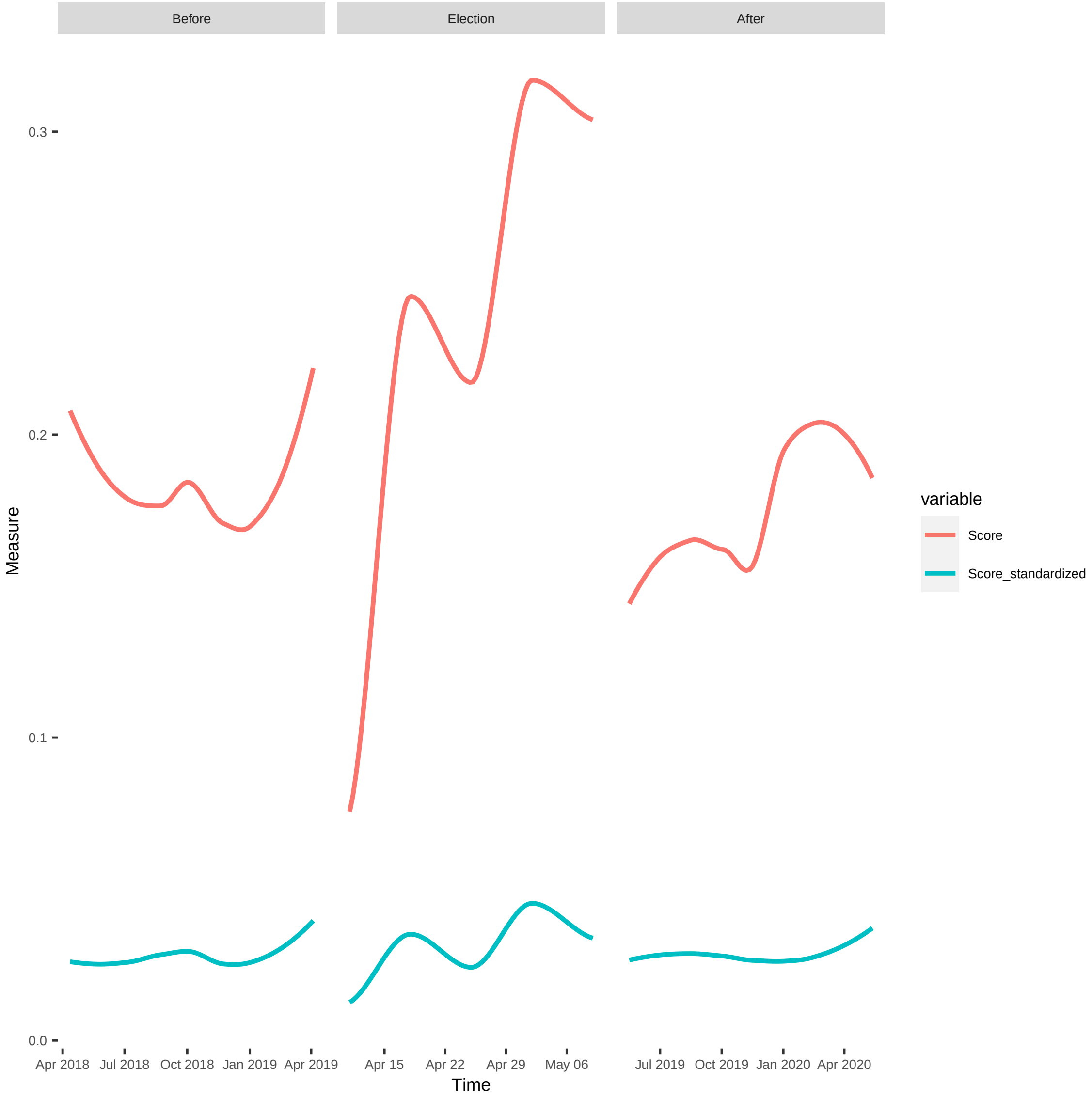
Notes: The figure inverse of the euclidean distance between negative tweets containing hashtags between BJP and INC politicians in English tweets. A higher value implies low polarization whereas a lower value implies higher polarization.The Before facet shows the pattern before the election, the Election period shows the pattern during the election whereas the After facet shows the pattern after election.

Figure 2.6: Distance metrics over time between BJP and INC politicians



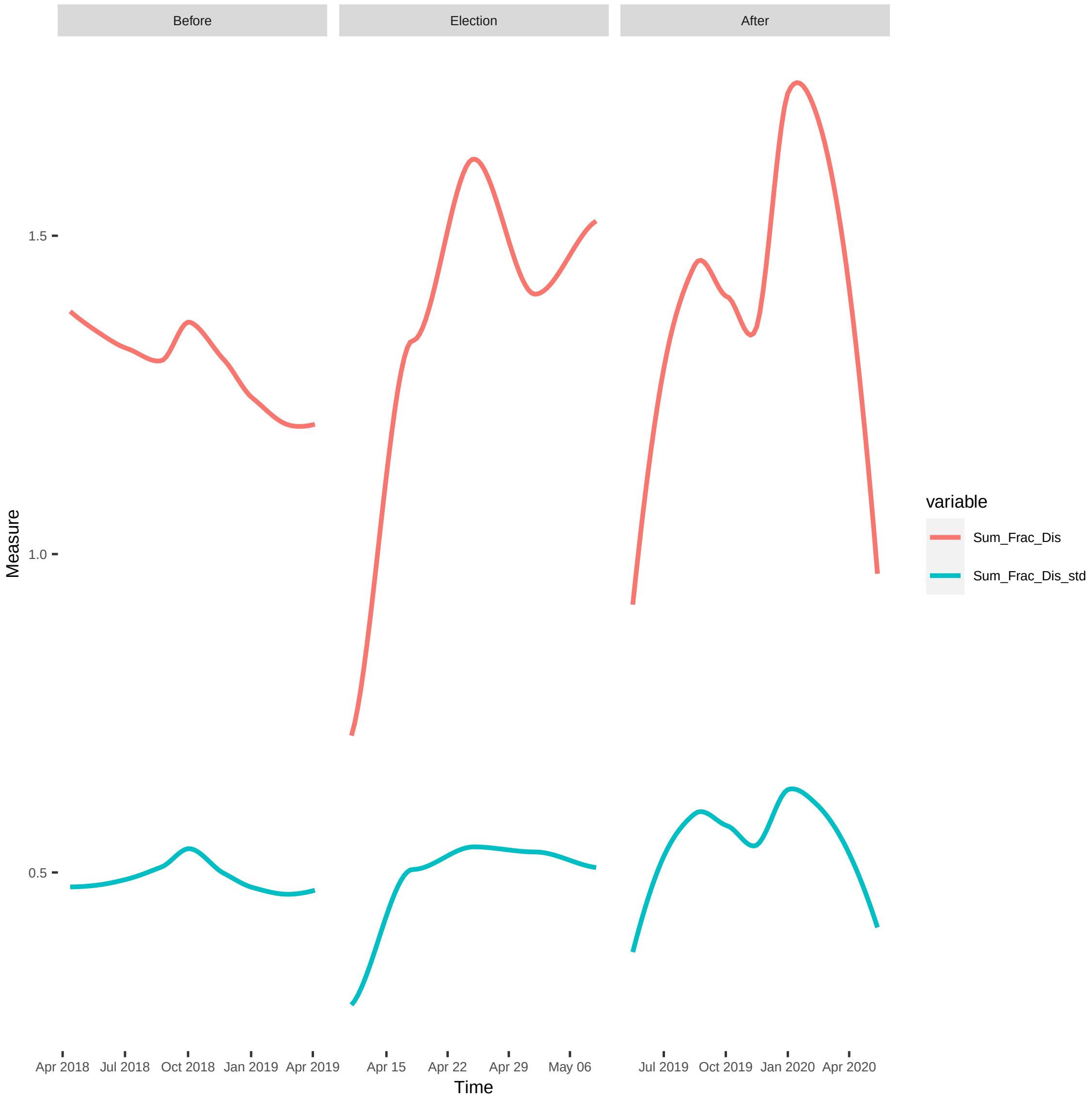
Notes: Euclidean distance between the probability distributions after applying the LDA model as measured by Hellinger, KLD DR, KLD RD. KLD refers to the Kullback Leibler Divergence. Distance between the words used and the list of words used as measured by JD DR and JD DLRL respectively. JD refers to the Jaccard distance. The Before facet shows the pattern before the election, the Election period shows the pattern during the election whereas the After facet shows the pattern after election.

Figure 2.7: Euclidean Distance between the dominant topics



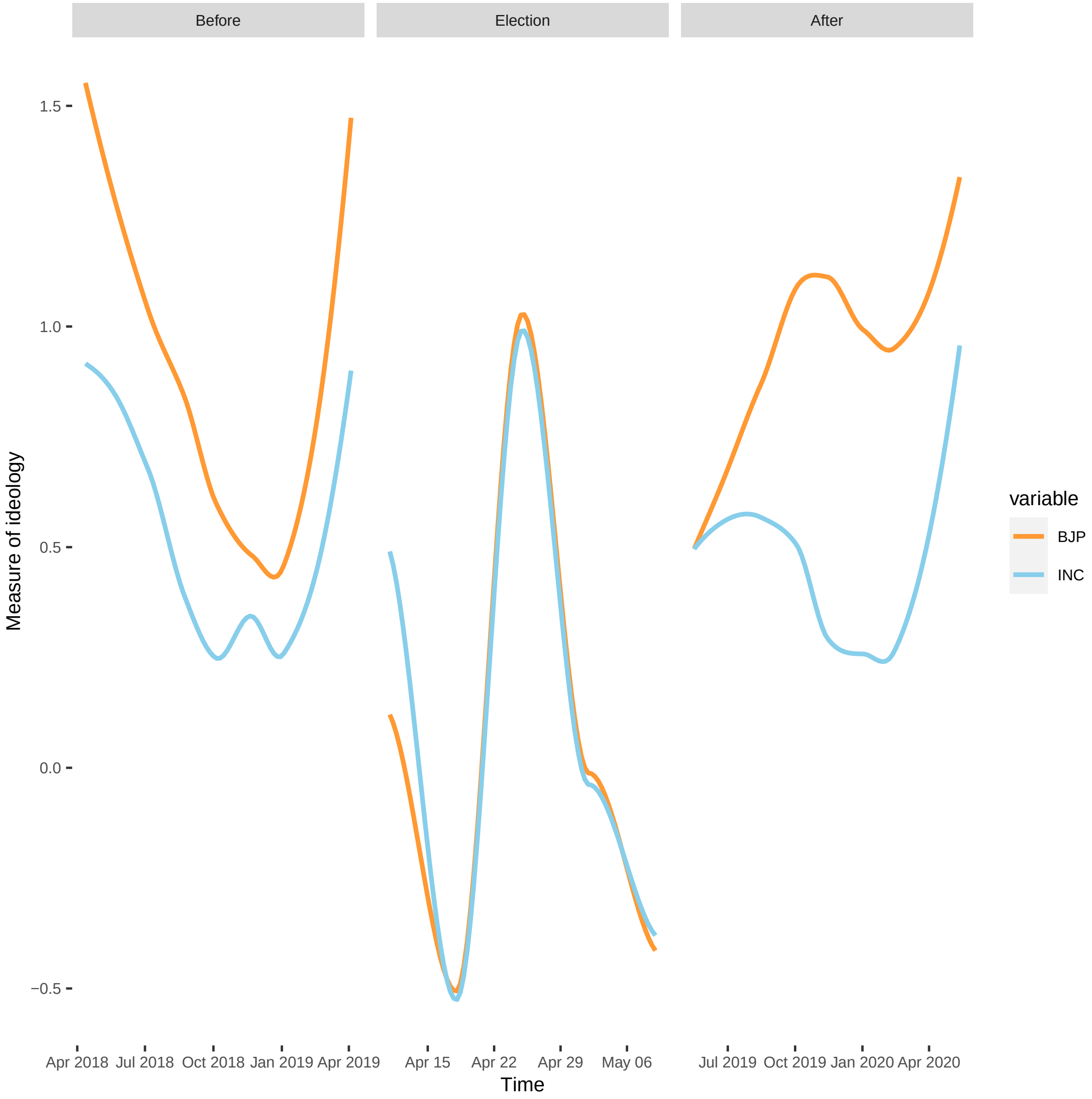
Notes: Score shows the euclidean distance between the vector of fractions of the tweets by the BJP and INC politicians devoted to the dominant topic calculate at at a weekly basis whereas Score_standardized shows the standardized version as explained in the text. The Before facet shows the pattern before the election, the Election period shows the pattern during the election whereas the After facet shows the pattern after election.

Figure 2.8: Euclidean Distance for fraction of positive, negative and neutral sentiments in a topic between the BJP and INC politicians



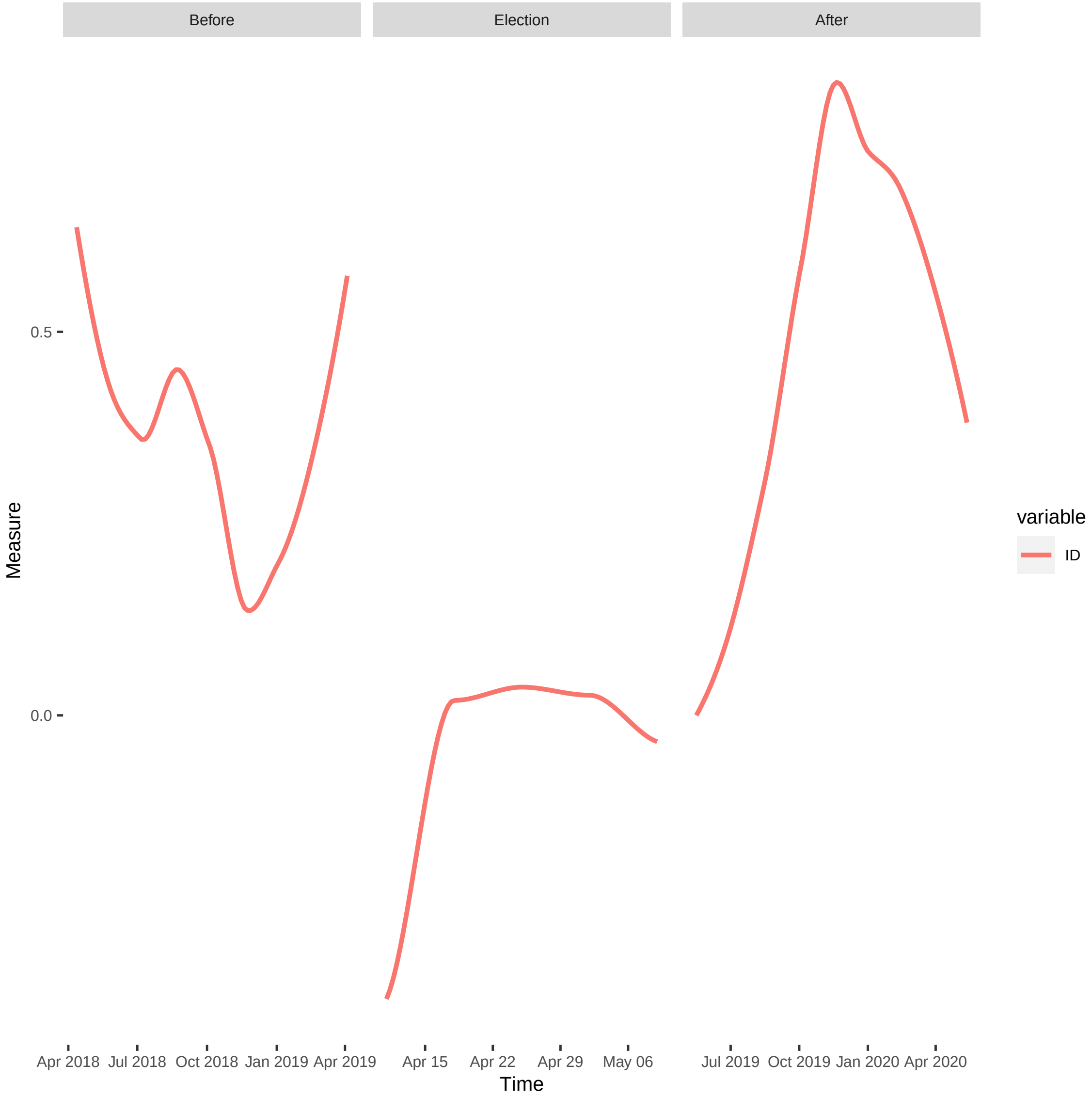
Notes: Sum_Frac_Dis measures the euclidean distance between the vector of fractions of tweets by BJP and INC politicians used with a positive, negative and neutral sentiments at a weekly basis whereas $Sum\_Frac\_Dis_{std}$ shows the standardized version as explained in the text. The Before facet shows the pattern before the election, the Election period shows the pattern during the election whereas the After facet shows the pattern after election.

Figure 2.9: Mean ideological score over time computed using Bayesian Ideal Point Estimation



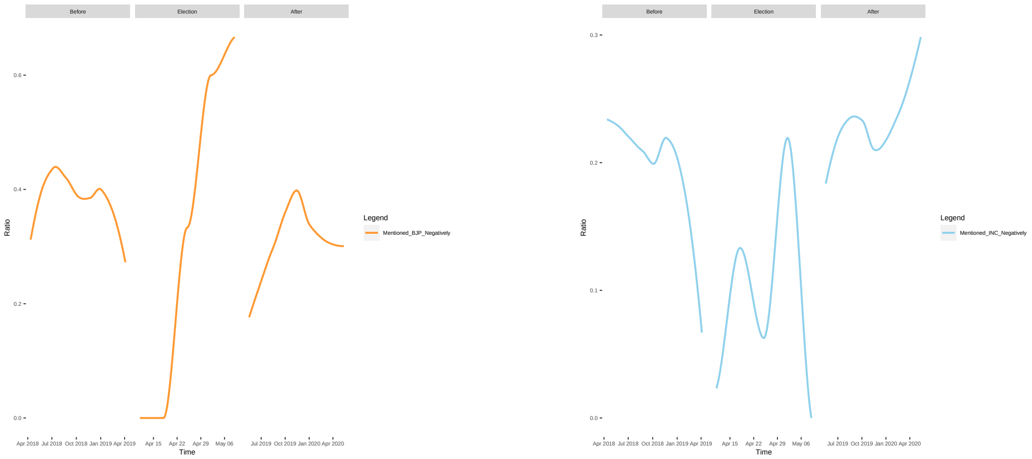
Notes: BJP and INC shows the movement of ideological scores for BJP and INC politicians respectively calculated using Bayesian Ideal Point estimation. The Before facet shows the pattern before the election, the Election period shows the pattern during the election whereas the After facet shows the pattern after election.

Figure 2.10: Mean ideological polarization over time computed using Bayesian Ideal Point Estimation.



Notes: ID is the difference between the mean ideological scores of of BJP and INC politicians. The Before facet shows the pattern before the election, the Election period shows the pattern during the election whereas the After facet shows the pattern after election.

Figure 2.11: Trend of negative mentions by parties over time



(a) Trend of negative mentions of BJP politicians by INC politicians

(b) Trend of negative mentions of BJP politicians by INC politicians

Notes: The Before facet shows the pattern before the election, the Election period shows the pattern during the election whereas the After facet shows the pattern after election.

Table 2.1: Number of total seats won by different parties over from 1984- 2019, at the national level

State	Total seats won by BJP	Total seats won by BSP	Total seats won by CPI	Total seats won by CPI(M)	Total seats won by INC	Total seats won by NCP
Andaman & Nicobar Islands	3	0	0	0	7	0
Andhra Pradesh	15	0	7	4	174	0
Arunachal Pradesh	5	0	0	0	11	0
Assam	29	0	1	2	60	0
Bihar	132	0	17	3	75	1
Chandigarh	4	0	0	0	5	0
Chhattisgarh	40	0	0	0	6	0
Dadar & Nagar Haveli	3	0	0	0	2	0
Daman & Diu	5	0	0	0	3	0
Delhi	43	0	0	0	28	0
Goa	7	0	0	0	9	0
Gujarat	175	0	0	0	79	0
Haryana	28	1	0	0	48	0
Himachal Pradesh	25	0	0	0	17	0
Jammu Kashmir	11	0	0	0	15	0
Jharkhand	32	0	1	0	8	0
Karnataka	109	0	0	0	136	0
Kerala	0	0	9	48	99	0
Lakshwadeep	0	0	0	0	7	2
Madhya Pradesh	227	4	0	0	128	0
Maharashtra	122	0	1	1	205	33
Manipur	1	0	1	0	12	1
Meghalaya	0	0	0	0	15	3
Mizoram	0	0	0	0	5	0
Nagaland	0	0	0	0	5	0
Odisha	32	0	3	2	70	0
Pondicherry	0	0	0	0	7	0
Punjab	13	5	1	0	53	0
Rajasthan	134	0	0	1	106	0
Sikkim	0	0	0	0	0	0
Tamil Nadu	8	0	9	6	109	0
Telangana	5	0	0	0	17	0
Tripura	2	0	1	14	4	0
Uttar Pradesh	350	82	3	1	157	0
Uttarakhand	14	0	0	0	7	0
West Bengal	24	0	25	180	58	0

Table 2.2: Regression Discontinuity Estimates for Hashtag Analysis

Panel A: RDD Estimates for Hashtag Similarity

	(1)	(2)	(3)	(4)	(5)
Degree	Hashtag_{10}	Hashtag_{20}	Hashtag_{40}	Hashtag_{50}	Hashtag_{100}
1	-1.149*	0.585	2.024	3.179*	-0.712
	(0.661)	(1.013)	(1.297)	(1.782)	(2.955)
2	-1.476*	-0.337	0.856	1.930	-0.883
	(0.758)	(1.308)	(1.679)	(2.070)	(3.364)
3	-1.565*	0.501	1.827	2.771	-1.247
	(0.847)	(1.579)	(2.311)	(2.519)	(3.877)
4	0.052	4.449	1.970	4.206	3.881
	(1.346)	(3.109)	(4.091)	(4.413)	(7.803)
N	104	104	104	104	104

Panel B: RDD Estimates for $Inverse\_Score\_std$ **estimates**

	(1)	(2)	(3)	(4)	(5)
Degree	$\text{Inv\_Score\_std}_{10}$	$\text{Inv\_Score\_std}_{20}$	$\text{Inv\_Score\_std}_{40}$	$\text{Inv\_Score\_std}_{50}$	$\text{Inv\_Score\_std}_{100}$
1	-80.796	-61.676	-85.921	0.452	-0.134
	(60.602)	(53.497)	(83.394)	(1.826)	(1.590)
2	-99.342	-100.640	-104.111	0.642	-1.098
	(71.024)	(74.989)	(100.885)	(2.781)	(2.029)
3	-110.202	-115.865	-113.233	-0.598	-1.000
	(78.748)	(84.178)	(110.437)	(3.556)	(2.344)
4	-110.112	-110.459	-133.120	-0.656	2.041
	(88.487)	(91.137)	(132.871)	(4.105)	(2.620)
N	104	104	104	104	104

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2] Hashtag similarity is defined as the number of common hashtags used by BJP and INC conditional on the top hashtags used by them. $Inverse\_Score\_std$ is the inverse of distance between fraction of negative tweets for the BJP and INC out of all the tweets that use a similar hashtag standardized for the number of topics. I perform a non-parametric RDD in Time using a data driven bandwidth selection method. The estimates are reported for 1, 2, 3 and 4 degrees of polynomials.

Table 2.3: Regression Discontinuity Estimates for Euclidean distance measures between topic distributions and word distributions

	(1)	(2)	(3)	(4)	(5)
Degree	Hellinger	KLD$_{BI}$	KLD$_{IB}$	JD$_{IB}$	JD$_{BLIL}$
1	-0.032	-0.036	-0.035	-0.067***	-0.071***
	(0.031)	(0.031)	(0.031)	(0.013)	(0.017)
2	-0.022	-0.014	-0.014	-0.070***	-0.080***
	(0.040)	(0.039)	(0.038)	(0.018)	(0.022)
3	-0.016	0.002	0.003	-0.090***	-0.091***
	(0.041)	(0.045)	(0.045)	(0.019)	(0.024)
4	-0.028	0.002	0.006	-0.049	-0.024
	(0.066)	(0.059)	(0.055)	(0.035)	(0.047)
N	104	104	104	104	104

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2] *Hellinger* Distance and the KLD_{BI} and KLD_{IB} measure the distance between the topic distributions after implementing the LDA model on the content of the tweets. The JD_{IB} and JD_{BLIL} measures the distance between the word distributions used by the Democrats and Republicans in their tweets. I perform a non-parametric RDD in Time using a data driven bandwidth selection method. The estimates are reported for 1, 2, 3 and 4 degrees of polynomials.

Table 2.4: Regression Discontinuity Estimates for Euclidean Distance and Euclidean distance interacted with sentiment for dominant topic

Panel A: RDD Estimates for Euclidean Distance

	(1)	
Degree	Score	Score_{std}
1	-0.063	-0.010
	(0.050)	0.015
2	0.023	0.007
	(0.046)	(0.021)
3	0.049	0.022
	(0.050)	(0.024)
4	0.001	0.035
	(0.100)	(0.033)
N	104	104

Panel B: RDD Estimates for sentiment augmented Euclidean distance

	(1)	
Degree	Sum_Frac_Dis	$\text{Sum\_Frac\_Dis}_{std}$
1	-0.176	-0.053
	(0.276)	(0.096)
2	-0.222	0.033
	(0.318)	(0.130)
3	-0.128	0.010
	(0.358)	(0.133)
4	-0.054	0.140
	(0.640)	(0.196)
N	104	104

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2] *Score* is defined as the euclidean distance between the vectors of fraction of each topic(dominant topic in each tweet) for the BJP and INC tweets for each week.*Sum_Frac_Dis* is the euclidean distance for each topic between BJP and INC, after controlling for the sentiment of the topics.$Score_{std}$ and $Sum\_Frac\_Dis_{std}$ represents the standardized version of the metrics, where the metric is divided by the square root of number of topics. I perform a non-parametric RDD in Time using a data driven bandwidth selection method. The estimates are reported for 1, 2, 3 and 4 degrees of polynomials.

Table 2.5: Regression Discontinuity Estimates for Bayesian Ideal Point(BIP) Estimates for politicians and Ideological Difference

Panel A: RDD Estimates for BIP Estimates for politicians

	(1)	(2)
Degree	BJP	INC
1	-1.802***	-0.820
	(0.549)	(0.450)
2	-2.305***	-1.195
	(0.778)	(0.814)
3	-2.409***	-1.301
	(0.934)	(0.933)
4	-2.576**	-1.434
	(1.035)	(1.088)
N	104	104

Panel B: RDD Estimates for Ideological Difference

	(1)
Degree	Ideological_Diff
1	-0.718**
	(0.337)
2	-0.768*
	(0.441)
3	-0.874
	(0.539)
4	-1.265*
	(0.666)
N	104

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2] Footnote 2: In **Panel A** BJP and INC shows the mean ideological score for all BJP and INC politicians. In **Panel B** $Ideological\_Diff$ difference between mean BJP and INC ideological scores. I perform a non-parametric RDD in Time using a data driven bandwidth selection method. The estimates are reported for 1, 2, 3 and 4 degrees of polynomials.

Table 2.6: Regression Discontinuity Estimates for Mention Network Analysis

Panel A: RDD Estimates for Mentions by INC

Degree	(1) BJP_Negatively
1	0.030
	(0.240)
2	0.021
	(0.308)
3	-0.433
	(0.492)
4	-0.545
	(0.090)
N	104

Panel B: RDD Estimates of mentions by BJP

Degree	(1) INC_Negatively
1	0.108
	(0.103)
2	0.114
	(0.123)
3	0.159
	(0.146)
4	0.194
	(0.179)
N	104

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.
[2] In **Panel A** *BJP_Negatively* are shares of how many times BJP politicians are mentioned negatively relative to all mentions of BJP politicians by the INC politicians. In **Panel B** *INC_Negatively* are shares of how many times INC politicians are mentioned negatively relative to all mentions of INC politicians by the BJP politicians. I perform a non-parametric RDD in Time using a data driven bandwidth selection method. The estimates are reported for 1, 2, 3 and 4 degrees of polynomials.

Table 2.7: OLS Sub-sample Estimates for Hashtag Analysis

Panel A: Estimates for Hashtag Similarity

Metric	(1) Before	(2) After
Hashtag$_{10}$	0.07143	0.2857**
	(0.14483)	(0.1010)
Hashtag$_{20}$	0.1667	0.04762
	(0.2546)	(0.23490)
Hashtag$_{40}$	0.1548	-0.09524
	(0.3978)	(0.21473)
Hashtag$_{50}$	2.463e-16	3.084e-16
	(4.859e-01)	(3.883e-01)
Hashtag$_{100}$	-0.0119	0.2500
	(0.9234)	(0.4597)
N	8	8

Panel B: Estimates of for*Inverse_Score_std* estimates

Metric	(1) Before	(2) After
Inv_Score_10$_{std}$	13.39	0.1503
	(10.93)	(0.6005)
Inv_Score_20$_{std}$	13.25	-1.0146
	(10.07)	(0.7933)
Inv_Score_40$_{std}$	17.21	-0.1608
	(14.80)	(0.8895)
Inv_Score_50$_{std}$	0.3028	-0.2333
	(0.3643)	(0.1790)
Inv_Score_100$_{std}$	0.2140	0.03625
	(0.3302)	(0.23343)
N	8	8

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2] Hashtag similarity is defined as the number of common hashtags used by BJP and INC politicians conditional on the top hashtags used by them. *Inverse_Score_std* is the inverse of distance between fraction of negative tweets for the BJP and INC politicians of all the tweets that use a similar hashtag standardized for the number of topics. The Before column shows the estimates for 8 weeks before the election whereas the After column shows the estimates for 8 weeks after the election.

Table 2.8: OLS Sub-sample estimates for Euclidean distances for topic distributions and word distributions

	(1)	(2)
Metric	Before	After
Hellinger	0.005685	-0.001612
	(0.007473)	(0.005830)
KLD_BI	0.004240	-0.0000642
	(0.008231)	(0.0037969)
KLD_IB	0.004188	0.0003346
	(0.008208)	(0.0038722)
JD_BI	0.004078	0.006455**
	(0.002211)	(0.002562)
JD_BLIL	0.003652	0.005875**
	(0.001990)	(0.002019)
N	8	8

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.
[2] $Hellinger$ Distance and the KLD_{BI} and KLD_{IB} measure the distance between the topic distributions after implementing the LDA model on the content of the tweets. The JD_{BI} and JD_{BLIL} measures the distance between the word distributions used by the BJP and INC politicians in their tweets. The Before column shows the estimates for 8 weeks before the election whereas the After column shows the estimates for 8 weeks after the election.

Table 2.9: OLS Sub-sample Estimates for Euclidean Distance and Euclidean distance interacted with sentiment for dominant topic

Panel A: Estimates for Euclidean Distance

Metric	(1) Before	(2) After
Score	0.006067	-0.009646
	(0.015524)	(0.007107)
Score_{std}	8.135e-06	-0.0005667
	(4.673e-03)	(0.0020767)
N	8	8

Panel B: Estimates for sentiment augmented Euclidean distance

Metric	(1) Before	(2) After
Sum_Frac_Dis	-0.007249	0.007772
	(0.043724)	(0.114610)
$\text{Sum\_Frac\_Dis}_{std}$	-0.008339	0.000714
	(0.015973)	(0.032930)
N	8	8

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.

[2]Score is defined as the euclidean distance between the vectors of fraction of each topic(dominant topic in each tweet) for the BJP and INC politicians tweets for each week. Sum_Frac_Dis is the euclidean distance for each topic between Republicans and Democrats after controlling for the sentiment of the topics. The Before column shows the estimates for 8 weeks before the election whereas the After column shows the estimates for 8 weeks after the election.

Table 2.10: OLS Sub-sample Estimates for Bayesian Ideal Point(BIP) Estimates for politicians and Ideological Difference in competitive and non-competitive districts

Panel A: Estimates for BIP Estimates for politicians		
	(1)	(2)
Metric	Before	After
BJP	0.30638***	0.10023
	(0.04509)	(0.08288)
INC	0.22778**	0.01242
	(0.08993)	(0.10145)
N	8	8

Panel B: Estimates for Ideological Difference		
	(1)	(2)
Metric	Before	After
ID	0.07860	0.08781
	(0.07352)	(0.06465)
N	8	8

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.
[2] In **Panel A** BJP and INC shows the mean ideological score for all BJP and INC politicians respectively. In **Panel B** ID and shows the difference between mean BJP and INC politicians' ideological scores. The Before column shows the estimates for 8 weeks before the election whereas the After column shows the estimates for 8 weeks after the election.

Table 2.11: OLS Sub-sample Estimates for Mention Network Analysis

Panel A: Estimates for Mentions by INC		
	(1)	(2)
Metric	Before	After
BJP_Negatively	-0.03842	0.05171
	(0.03489)	(0.03195)
N	8	8

Panel B: Estimates of mentions by Republicans		
	(1)	(2)
Metric	Before	After
INC_Negatively	-0.02223	0.00169
	(0.01200)	(0.02322)
N	8	8

[1] Standard error in parentheses. * denotes 10 percent significance, ** denotes 5 percent significance and *** denotes 1 percent significance.
[2] In **Panel A** *BJP_Negatively* are shares of how many times BJP politicians are mentioned negatively relative to all mentions of BJP politicians by the INC politicians. In **Panel B** *INC_Negatively* are shares of how many times INC politicians are mentioned negatively relative to all mentions of INC politicians by the BJP politicians. The Before column shows the estimates for 8 weeks before the election whereas the After column shows the estimates for 8 weeks after the election.

Chapter 3

The effect of political business cycles on government expenditure and night light in India (with Sagnik Das)

3.1 Introduction

Incumbent politicians are believed to manipulate voters close to an election to maximize their chances of getting re-elected. This leads to political business cycles, wherein politicians stimulate the economy close to the election. In this chapter, we investigate if there is evidence of political business cycles in the Indian context. We investigate if government expenditure increases close to the election.

We use labour expenditure for employing manual labourers, employment provided to labourers, new road constructed data as measures of government expenditure and night light data at the yearly level as proxy for electricity provision as well as some other government expenditure induced development. We also use high-frequency night light data at the monthly level to check for political business cycles over the short run. We find the relationship between these variables and the proximity to an election to investigate the existence of political business cycles after controlling for economic, demographic and other election-related controls along with spatial and temporal fixed effects.

To get the data on elections, we focus on state-level legislative assembly elections for different states of Indian held broadly within the period 1993-2018 with slight adjustments as per the period of availability of the outcome variable. Our outcome variables, however, are not at the assembly constituency level but the district level. Since a district comprises many assembly constituencies, the election data are aggregated at the district level. States comprise districts, and all constituencies in a state have elections in the same year. Hence, it would be possible to conduct the analysis at the state level too. However, there is considerable heterogeneity among the districts within a state and between districts belonging to different states. To capture this heterogeneity across districts, we keep our analysis at the district level. Since political competition among the contesting political parties participating in the election might be the underlying reason that gives rise to political business cycles, we control for political competition in each district by computing the difference in the share of the number of legislative assembly constituencies that the electorally most successful and the second most successful political party won in a district for each election cycle.

Our results indicate the existence of political business cycles for variables measuring labour expenditure made by the government and employment provided at the intensive margin. We also find evidence of political business cycles impacting the length of new road constructed and the cost sanctioned for new roads to be constructed. We find some evidence of political business cycles existing for night light data at the yearly level. However, we do not find a conclusive trend for the high-frequency analysis at the monthly level, even though we find a spike in night light intensity one month before the election. There is no evidence of gaming the electrification process close to the election.

The theory of political business cycles was first proposed by Nordhaus (1975) and Lindbeck (1976). While these theories focus on myopic voters who focus on outcomes close to the election to decide whom they are going to vote for, there is another strand of literature developed by Rogoff and Sibert (1988) and Marin et al. (1990) suggesting that stimulating the economy close to the election might provide a signal to voters about the ability of politicians to influence future policies. Whether one believes that voters are myopic or forward-looking, the theory suggests that politicians would be inclined to increase economic activity close to elections to gain electoral advantage.

The theory of political business cycles has been primarily tested in developed and industrialized nations for macroeconomic policies, and the evidence has been mixed. Berger and Woitek (1997) find evidence of political business cycles in Germany whereas McCallum (1978), and Klein (1993) reject the hypothesis that macroeconomic outcomes are influenced by the political business cycles. The evidence from developing nations in more micro-founded economic outcomes over which local politicians have more significant control is increasing over time. Gonzalez (2002) shows that the Mexican government uses public spending to secure votes. In the Indian context, Cole (2009) shows that public banks in India track the electoral cycle with agricultural credit growing up by 5 to 10 percent in the election year. Baskaran et al. (2015) shows that electricity provision as proxied by night lights fluctuates in accordance with electoral cycles; however, they only focus on by-elections in their paper where a special election is held upon the death of an incumbent politician.

Our chapter extends the political business cycle literature and makes three significant contributions. First, we exploit a host of developmental outcome variables with spatial and temporal variations across districts in India to study the impact of political business cycles. Our second contribution is that we look at new roads created data and the employment data from the NREGA programme, which has not been used in the political business cycle literature previously and which helps us understand whether the creation of infrastructural goods as well as employment outcomes are also influenced by political business cycles. Thirdly, we also look at short-run fluctuations in high-frequency night light data and investigate if they are influenced by election cycles.

The rest of the chapter is organized as follows. Section 3.2 discusses the Institutional Background and Section 3.3 introduces the Data. Section 3.4 discusses the Empirical Strategy, Section 3.5 discusses the Results and Section 3.6 concludes.

3.2 Institutional Background

India is a Unitary Federation divided into 29 states (which comprise almost the entire area and population) and seven union territories. All 29 states and 2 Union Territories (the National Capital Territory of Delhi and Puducherry) have a parliamentary form of provincial government. There is a parliamentary form of government at the Centre, known as the central government for the entire country and also at the State level. The Constitution of India vests significant executive power to state governments, whose combined expenditure outstrips that of the central government.

3.2.1 Geographical hierarchy of India according to the ECI

India's hierarchical structure is three-tiered, as implemented by the Election Commission of India (the non-political central body responsible for conducting elections in India). The Central Government, known as the Lok Sabha, is at the top of the tier. The Lok Sabha has 552 members, out of which 530 members representing the different states and 20 members representing the different union territories are elected through the Lok Sabha elections conducted every five years. The party or the coalition that secures the majority of votes forms the executive branch comprising the Prime Minister (P.M.), the Cabinet and the Council of Ministers. 2 members of parliament are assigned through the President's (the nominal head of the country) recommendation. The Lok Sabha is mainly concerned with undertaking national-level policies as defined in the Union List and Concurrent List, which mainly comprises defence, foreign affairs, railways, banking, education and others. On the second tier of the three-tier system lies the state legislative assemblies, also known as the Vidhan Sabhas. Each state is divided into single-member Legislative Assembly constituencies for which elections are held with the winner decided by a first-past-the-post (FPTP) system. Elections are held regularly at five-year intervals (elections can be held before the end of a five-year term if no party or coalition can continue with a majority). Elections are conducted by the Election Commission of India, a constitutionally established independent body. The party or coalition that secures the majority of Legislative Constituency seats forms the executive branch comprising the Chief Minister of the state, the Cabinet, and the Council of Ministers. The electorate votes partly based on the ruling political party and its chief ministerial candidate and partly based on the legislative constituency candidate (as state election survey results in Lokniti-CSDS [2014b] show. However, the former's importance is generally higher than that of the latter; the importance attached to the legislative constituency candidate can be over 30 percent, as in the 2018 Karnataka Legislative Assembly election).

The number of Assembly Constituencies into which a state is divided varies markedly across states: the most populated state, Uttar Pradesh, has 403 assembly constituencies while Puducherry has only 30. Among the more populated states, the number of legislative constituencies is roughly proportional to the state's total population according to the 1971 census. For states with lower populations, the population-constituency ratio is higher.

At the last tier of the three-tier political and administrative system is a local government system called the Panchayat (Rural Local Bodies) and Municipalities (Urban Local Bodies). The Panchayat or Municipalities is a local government system comprising elected members through panchayat/municipal elections, conducted every five years. The panchayat system is three-tiered, with the village council at the very bottom, followed by the block council and the district council at the very top.

Although the central government and the state government function independently, when it comes to providing public goods or implementing specific policies decided upon by both the state government and the central government, many decisions are taken at the constituency level. Argued by Lehne et al. (2018), even though the members of the legislative assembly do not have a direct influence on the process of provision of public goods sanctioned by the central government, the members of the legislative assembly enjoy substantial political power to have an influence on the implementation of a policy or the provision of public goods at the local level. Also, the members of a legislative assembly constituency have direct access to the Chief Minister's office, which is the state's highest administrative office. Thus, the legislative assembly member has the potential to inform the ministers and the Chief Minister of the state about the developmental needs of his/her constituency leading to influence infrastructural or any other socio-economic changes in the constituency. Keeping in mind this potential of the legislative assembly members to influence government expenditure in the assembly constituencies, in this chapter, we investigate if the influence is large enough to bring about significant changes in their localities, resulting in political business cycles.

3.2.2 Geographical hierarchy of India according to the Census

The Census of India divides India's geographical boundary into States, and within those states are districts, which are further divided into blocks and then into towns and villages. A district also constitutes several legislative assembly constituencies. The population census is conducted by the Government of India every ten years. Since Indian independence, there have been many changes in these boundaries, mentioned in the different census reports across the years. New states have come into being within our period of study, and also, the geographical boundaries of districts have been redrawn multiple times. Since our study period overlaps across three different population census periods (1991, 2001 and 2011), to avoid any potential overlaps in boundaries across different census periods, we have considered districts' geographical boundaries as drawn in the 2011 population census in our analysis. The outcome variables used in the chapter are compatible with the population census 2011 delimitation of district boundaries.

3.3 Data

3.3.1 Outcome Variables

For our outcome variables, we use three data sources - employment data from Mahatma Gandhi National Rural Employment Guarantee Act (NREGA), new road constructed data from Pradhan Mantri Gram Sadak Yojana (PMGSY) and night lights data both at the yearly level and at the monthly level (for the high-frequency night light analysis). Our outcome variables are at the village level, and we aggregate them to the district using a district identifier as per the 2011 delimitation reports of population Census[1].

3.3.1.1 First Outcome Variable: Mahatma Gandhi National Rural Employment Guarantee Act

For our first outcome variable, we use data from one of the most extensive employment guarantee programmes in the world implemented in India. This employment guarantee programme started in 2006 with the enforcement of the National Rural Employment Guarantee Act (NREGA) in 200 of India's most backward districts, Ambasta et al. (2008). It was then extended to all of rural India from April 1,

[1] The high-frequency night light data and the NREGA data is already at the district level and hence we do not need to aggregate it

2008. This is a unique act because it legally binds the government to provide work to rural labourers who want to work at that wage, and the government cannot be excused due to a lack of resources. In this scheme, the government is liable to provide 100 days of employment to every rural household whose adult members volunteer to participate in unskilled manual labour. The employment seeker has to register himself/herself through the Gram Panchayat and is given a job card. The employment begins within 15 days of the issuance of a job card. If the government cannot provide employment within those 15 days, they need to pay the labourers' unemployment benefits. Some other salient features of the scheme are that one-third of the beneficiaries need to be women. Some form of employment-related fundamental rights such as access to drinking water, restroom, creches for kids, emergency health care need to be ascertained. The scheme also needs to be socially audited, but the audit process has come under criticism for not being efficient and effective.

The scheme is implemented at the national, state, district and village level, and all the tiers have their specific role to play. Although the funding decisions are mostly made at the Central level, and the planning and administration decisions are made at the State level, the actual issuance of job cards after proper scrutiny and allotting of jobs happen at the Gram Panchayat level.

We have the NREGA data available from 2011 to 2020. Using the NREGA dataset, we construct a host of outcome variables. The total number of households who demanded jobs, the share of households (HH) who received the job card out of all job card applicants, the share of households who were allotted jobs out of all those households who demanded jobs, the share of households who went onto to complete 100 days of work conditional upon receiving allotted jobs, and finally, the labour expenditure by the government. Table 3.1 shows the summary statistics for the variables. Table 3.2 shows these variables' values from four years before the election to the election year. We see some variation in the demand for jobs, although they do not show any particular trend. We can see that for the share of applicants who received job cards and the share of applicants who were allotted work, there is no change in the values over the years because these are determined by bureaucratic processes which are well defined. There is not much scope for political interference. We see considerable variation in the labour expenditure over the years, and the value mostly keeps increasing from four years before the election to the election year. The share of applicants who completed 100 days of employment also keeps increasing from four years before the election to one year before the election but drops in the election year. For each of the variables mentioned, we also compute the yearly change in each variable compared to the previous year. Figures 3.1, 3.2a and 3.2b show the spatial and temporal variation for the total number of households who demanded job, share of households who were allocated job cards and subsequently allotted jobs. There is some spatial and temporal variation in the total number of households who demanded work. However, we see minimal spatial and temporal variation in the share of households with issued job cards and the share of households with allotted jobs. There is minimal variation in these variables, given that they are not heavily influenced by elections and are determined by purely manual bureaucratic procedures. Figures 3.3a and 3.3b shows the spatial and temporal variation in the labor expenditure and share of households who completed 100 days of unemployment. We see substantial variation both across districts and distance to an election in these two variables.

3.3.1.2 Second outcome variable : Pradhan Mantri Gram Sadak Yojana

In 2000, India launched the Pradhan Mantri Gram Sadak Yojana (Prime Minister's Village Road Program, or PMGSY). This national programme aimed to build a paved road to every village in India eventually. While the federal government issued implementation guidelines, decisions on village-level allocations of roads were ultimately made at the district level. The unit of targeting road construction was the habitation, the smallest rural administrative unit in India. A village typically comprises between one and three habitations; there are approximately 600,000 villages in India and 1.5 million habitations. In some states, this took the form of a strict population threshold for road construction eligibility, while other states used other criteria. Given the programme rules, early-treated villages tended to have larger populations but were not substantially different from late-treated villages in other characteristics. There were initially 80,000 villages eligible for the road construction program, which has grown as guidelines have been expanded to include smaller villages. By 2015, over 115,000 villages had paved roads built or upgraded under the PMGSY program. These construction projects were most often managed through subcontracts with larger firms and were built with capital-intensive methods and external labour. The PMGSY data is available from 2000 to 2014.

From the SHRUG ancillary data used previously in Asher and Novosad (2020), we use the information on the length of new road constructed and the costs sanctioned for new road to be constructed in each district in each year over our period of study and use those as our outcome variables. Table 3.1 shows the summary statistics for the new road constructed data and the cost sanctioned for new road to be constructed. We also compute the change in the variables. Table 3.2 shows the summary statistics for the same outcome variables based on distance to election. We see that for the length of new road constructed increases from four years before the election to two years before the election and then decreases in the year before the election and the election year. The variable on cost sanctioned for road to be constructed increases in value from four years before election to the year before the election and then falls in the election year. Figure 3.4a and Figure 3.4b show the spatial and temporal heterogeneity. While we find some spatial heterogeneity across states, the temporal variation seems to be limited to certain states only.

3.3.1.3 Third Outcome Variable : Night Light Data

For the third outcome, we use the night lights data from the SHRUG data set used previously in Asher et al. (2020). Night lights are a proxy for economic growth that has the advantage of high resolution and objective measurement over 20+ years (Henderson, Storeygard and Weill 2011). Night light data has been used in various papers in India because of the absence of official data on electrification or economic performance at regular intervals, especially for rural areas. Dugoua et al. (2018) use night light data to measure electrification in rural areas in India, Asher and Novosad (2017) use night light data to measure economic output in India.

As a measure of the intensity of night-light data, we use the district mean of calibrated night-light intensity. Table 3.1 shows the summary statistics for the night light data. Table 3.2 shows the summary statistics for night light data based on distance to election. The value of mean calibrated night light intensity increases from four years before the election to the election, although modestly. We also compute the changes in these variables. Figure 3.6 shows that we have significant heterogeneity in the spatial dimension, but not so much in the temporal dimension. Our yearly night light data range from 1994 to 2014.

To check for whether there is a more short-run effect of an election, we look at high-frequency night light data available at the monthly level. We use the India Lights platform, a repository of light output at night for 20 years for 600,000 villages across India[2].

Variables in the high frequency night light dataset The variables that we have in the dataset include the year and the month in which the measurement was taken, the name of the district and the state corresponding to that particular measurement, the satellite used to make the observation, the number of measurements in that month and the median night light intensity in that month in that year in that district[3]. Table 3.3 and Table 3.4 shows the summary statistics for the median night light intensity and the standardized median night light intensity respectively. As we can see, there is not much variation in the mean of the median night light intensity at various points in time. Figure 3.7 and Figure 3.8 shows the spatial and temporal variation in the median night light intensity and the standardized median night light intensity respectively. We can again see that the spatial component variation is more than the

[2] The Defense Meteorological Satellite Program (DMSP) has taken pictures of the Earth every night from 1993 to 2018. In collaboration with the World Bank, researchers at the University of Michigan used the DMSP images to extract the data on Indian night lights.

[3] The DMSP raster images have a resolution of 30 arc-seconds, equal to roughly 1 square kilometre at the equator. For each raster image, a pixel of the image is assigned a number based on a relative scale which ranges from 0 to 63. The number 0 indicates no light output, while 63 indicates the highest level of output. This assigned number is relative and might change depending on the satellite's sensor's gain settings, which constantly adjust to current conditions as it takes pictures throughout the day and at night. To derive a single measurement, the light output values are extracted from the raster image for each date for the pixels that correspond to each village's approximate latitude and longitude coordinates. Then the data is processed through a series of filtering and aggregation procedures. After extracting the data from the pixels, in the first step, according to recommendations from the National Oceanic and Atmospheric Administration (NOAA), data with too much cloud cover and solar glare is filtered out. Then, the resulting 4.4 billion data points are aggregated by taking the median measurement for each village over a month, adjusting for differences among satellites using multiple regression on year and satellite to isolate each satellite's effect. The median village light output within each administrative boundary for each month in the twenty years is determined for analysing the data at the district or state level. These monthly aggregates for each village, district, and state have been made accessible through the API.

variation in the temporal component. The monthly night light data set is available for the years 1993 to 2013.

3.3.2 Independent Variables

The independent/control variables include election-related variables, demographic and economic characteristics. The election-related variables have been constructed from the Trivedi election dataset for assembly constituency elections spanning 1989-2018 in the data repository of Lok Dhaba Election Data of Ashoka University. It is a dummy variable indicating how far is the election in a district from the year in which the outcome variable is measured. For example, the district of Bardhaman in West Bengal, a state in India, had its election in 1991 and 1996. We consider the election year of 1996 to be our baseline. Any outcome variable measured in 1995 is considered to be one year away from the election and is assigned a dummy. This is done for all the years in between till we reach the preceding election year of 1991, and the process is repeated. The distance to the election is our primary variable of interest. The dataset on election results reports other election details for each constituency in each district in each state for each election cycle. Since our level of analysis is at the district level, we aggregate the constituency-level variables like the share of seats in a district won by the most and the second most electorally successful party in a district for each election cycle and then take their difference. We use this as a control variable as a proxy for the district's electoral competitiveness. We also use the information on the majority party to be able to control for party fixed effects.

For the demographic variables, we have used the Population Census of 1991, 2001 and 2011. Using the variables reported at the district level for each of the censuses, we have controlled for population and some essential demographic variables such as sex-ratio, share of the literate population out of total population, share of literate population who completed primary schooling, share of literate population who completed secondary schooling and share of literate population who completed higher secondary schooling.

We use rural and urban employment figures provided by the Reserve Bank of India for the economic variables. For the analysis using the NREGA dataset, we also use the State Gross Domestic Product (SGDP) as a control variable. We cannot use SGDP for the other datasets because we do not have SGDP data for those periods.

3.4 Empirical Strategy

3.4.1 Analysis based on years to election

Our study's objective is to investigate the existence of a causal relationship between the proximity to the next election in a district and government expenditure in that particular district. In other words, we want to quantify the effect an upcoming election has on the government expenditure in the area through the actions of the elected representatives who want to maximize their probability of winning.

Ideally, our unit of analysis should be an assembly constituency because the jurisdiction of an elected representative's power is at the assembly constituency level. However, due to the non-availability of data, we have to aggregate the assembly constituency level electoral variables at the district level to match the spatial coarseness of the outcome variables as done in Clots-Figueras (2012).

We perform the analysis at the district using the following empirical strategy.

$$y_{dst} = \theta_d + \beta_t + \sum_{i=-4}^{-1} \gamma_i \times TOE_{st}^i + party_{dst} + \alpha \times SS_{dst} + Dem_{dt} + Eco_{st} + u_{dst} \quad (3.1)$$

where y_{dst} represents the outcome variable for a district d in state s in year t, θ_d represents district fixed effect, β_t represents the time fixed effect, TOE_{st} are dummies indicating years to election in state s at time t, the election year is treated as the baseline and γ_i is our variable of interest. $party_{dst}$ shows the electorally most successful party in district for each election cycle d, state s and time t and accounts for party fixed effect. SS_{dst} represents the seat share difference of the most and the second most electorally successful party in the district. Dem_{dt} and Eco_{st} are the demographic and economic controls at the district and the state level respectively.

3.4.2 Analysis based on high frequency night light data and months to election

To check if there is any effect of proximity of election on the short run night light intensity, we investigate the effect of the closeness of election on high-frequency night light intensity measured monthly. We merge each monthly observation of night light data to the election that happened before if the past election happened 30 months ago or two and a half years ago, and to the upcoming election if the next election is less than or equal to 30 months away or two and a half years away from the current month. We then drop the duplicate observations without loss of generality.

After we have the clean dataset, we create twenty-five dummies, 12 for each month upto one year before the election, one for the election month, and 12 dummies for each month upto one year after election. All other months are considered to be the baseline category. We also use district-month fixed effect to account for the seasonality in night light data in different districts and control for year fixed effect, along with all the other economic and demographic controls, seat share difference and party fixed effect. Our object of interest is to find what happens to night light as we move towards the election and what happens as we move away from it.

To answer the question we use the following specification

$$y_{dmst} = \theta_d \times \omega_m + \sum_{i=-12}^{12} \gamma_i \times 1[MTE_{smt}^{i} = 1] + \beta_t + party_{dst} + \alpha \times SS_{dst} + Dem_{dt} + Eco_{st} + u_{dst}, \tag{3.2}$$

where y_{dmst} is the night light in district d, in state s, in month m of year t, MTE_{smt}^{i} is a dummy variable that takes value 1 when the observation in state s in year t in month m is i months away from election. For example MTE_{smt}^{-1} is equal to 1 when the observation is one month before the election. $\theta_d \times \omega_m$ is the district times month fixed effect, β_t is the time fixed effect, $party_{dst}$ is the party fixed effect, SS_{dst} is the difference in seat share between the majority and the second majority party, Dem_{dt} are the demographic controls and Eco_{st} are the economic controls.

3.5 Results

3.5.1 Results for analysis based on years to election

In this section, we describe and interpret the regression equation results mentioned in the Empirical Strategy section for the yearly analysis, involving the various outcome variables mentioned in the Data section.

All the coefficient estimates of the primary explanatory variable in the chapter, Time to an election, viz. the distance from a state assembly election in years, are to be interpreted compared to the baseline category viz. year of the election.

3.5.1.1 Outcome variable:Mahatma Gandhi National Rural Employment Guarantee Scheme data

Table 3.5 shows the regression results when the outcome variable is the amount of money in 100,000 rupees spent by the government on labour expenditure. As referred to by the column names of the table, all the different model specifications indicate a trend in the effect of the proximity to election captured by the explanatory variable TOE_{st}. For the main model specification, represented in column (4), as we move closer to the election, the regression coefficient on TOE_{st} increases, indicating that, as we move closer to the election year, compared to the baseline year, even though the money spent on labour expenditure is less, it still shows a steady increase. As indicated by the regression coefficients, the coefficient for four, three and two years are statistically significant, but not for one year to election. However, the coefficients indicate a pattern, with the coefficient being 5.5 million rupees more than that in the election year for one year before the election and almost 74 million rupees less than that in the year of election for four years before the election. Also, the coefficient for the variable difference in seat share, which is a variable capturing the potential electoral competitiveness in a district, is negative and statistically significant. This implies that the higher the difference in the seat share between the most successful and the second most successful political party in a district in a particular election cycle, the less is the money disbursed by the government for labour expenditure in that district. This means that the less competitive the district is electorally, the lower is the money disbursed by the government for labour expenditure. Figure 3.9a plots the coefficient of TOE_{st}, and it is clear that there is an upward trend in the coefficient as the proximity to the election increases. This result suggests an effect of the political business cycle on money disbursed by the government for labour expenditure. We also computed the year-wise change in the money disbursed by the government for labour expenditure and then regressed it on TOE_{st} and other control variables. In the appendix, Table C.1 reports the results of this regression. We do not find any statistically significant effect there.

An increase in the labour expenditure can be caused by either an increase in the demand for NREGA jobs or an increase in the supply of jobs. A possible explanation for an increase in demand for jobs caused by the proximity to the election could be that, as an election approaches, the electorate expects that it would be possible to get more jobs by demanding more jobs from the local administration by using the promise of votes in exchange for NREGA jobs. If this mechanism exists, this effect will show up if we regress the number of households who demanded jobs on our main explanatory variable. Table 3.6 reports the regression result when the logarithm of the total number of households who demand work is regressed on TOE_{st} and other control variables. From column (4) of the table, it is observed that there is no statistically significant effect of the proximity to an election on the number of households demanding jobs. We also regress the logarithm of the change in the total number of households who demanded jobs on TOE_{st} and other control variables. The results are reported in Table C.2. We do not find a statistically significant effect or even trend following the proximity to an election.

To explore supply-side effects, we investigate if the share of job cards issued and the share of households allotted work have gone up to see if the main driver of the increase in labour expenditure is more work offered at the extensive margin. We regress the share of households who were issued job cards and the share of households who were allotted jobs to the number of households who demanded jobs on TOE_{st}, our primary variable of interest and other control variables. Since a job card being issued or a household being allotted work are bureaucratic processes, we should not expect to see these variables be affected by proximity to the election. Table 3.7 shows the regression result for the share of households issued job cards, and Table 3.8 shows the regression result for the share of households who were allotted jobs. From both the tables, it can be concluded that there is not much meaningful effect of proximity to the election. Therefore, it seems that the increase in labour expenditure are not driven by jobs on the extensive margin, which we were not expecting in the first place as these processes are generally immune to political influence. In the appendix, Table C.3 and Table C.4 reports the regression results when the change in the ratio of the number of households who were issued job cards to the number of households who demanded job and the change in the ratio of the number of households who were allotted jobs to the number of households who demanded jobs are regressed on TOE_{st} and other control variables. We do not find any statistically significant effect of TOE_{st} on either of the two outcome variables mentioned.

To check if the increase in money disbursed by the government for labour expenditure with an increase in proximity to the election can be explained by the increase in the supply of jobs at the intensive margin, we regress the number of households who completed 100 days of work conditional on the fact that these households were already allotted jobs. Table 3.9 reports the results of this regression. With reference to our main model specification results as shown in column (4) of the table, there is a statistically significant spike in the number of households completing 100 days of work conditional on the number of households who were allotted work, as the proximity to the election increases. Although the coefficient one year before the election is less than that two years before the election, there is a trend from four years before to one year before the election. Figure 3.9b shows this increasing trend in the political business cycle coefficient as it gets closer to the election year.

The results discussed above corroborate our theory that there is evidence of the political business cycle affecting work provided to the labourers. It also indicates that the possible mechanism through which developmental works might be affected by the political business cycle is through the proactive measures taken by the political parties who influence the functioning of the NREGA scheme at the local level.

3.5.1.2 Outcome variable: PMGSY data

Table 3.10 reports the regression results when the total district sum of the length of new road constructed under the PMGSY scheme is regressed on TOE_{st} and other controls. According to the main model specification shown in column (4) of the table, we see a consistent, statistically significant increase in the length of new road constructed under the PMGSY scheme with the increase in the proximity to the election. Compared to the election year, the year before the election reports that the length of the new road constructed under PMGSY was around 33 kilometre more. Figure 3.12a shows a clear upward trend in the regression coefficients of proximity to the election. The same regression for the change in total district sum of the length of the new road constructed under the PMGSY scheme is reported by Table C.6 in the Appendix. We do not find any statistically significant effect of the political business cycle on the change in total district sum of the length of new road constructed under the PMGSY scheme.

Table 3.11 reports the regression results when the total district sum of costs sanctioned for new road to be constructed under the PMGSY scheme is regressed on TOE_{st} and other controls. Like the total district sum of the new road's length, this variable also shows a steady increase compared to the baseline, as the proximity to the election increases. The coefficient is highest for one year before the election, as reported by column (4) of the table. According to column (4), compared to the year of the election, the previous year experiences more than 120 million rupees more sanctioned for the construction of new road. Figure 3.12b shows the steady increase in the coefficients as the proximity to the election increases, thus reflecting the impact of the political business cycle on cost sanctioned for road constructed under the PMGSY scheme. Cost sanctioned for new road constructed increases by a lot from 63 million rupees to 111 million rupees from three years before the election to two years before the election. This probably happens because the output of such sanctions in terms of new road constructed will be visible close to the election. Also, cost sanctioned is the highest just before the election year. The politicians probably use the announcements of cost sanctioned to lure voters for voting and also as a signal that the projects will be implemented if they are elected to power. Table C.7 in the appendix reports the results for the same regression, with the outcome variable being change in total district sum of costs sanctioned for new road to be constructed under the PMGSY scheme. Again, we do not find any statistically significant effect of the political business cycle on change in total district sum of costs sanctioned for new roads to be constructed under the PMGSY scheme.

3.5.1.3 Outcome variable: Night-light data

Table 3.12 reports the regression results when district mean of calibrated night-light intensity is regressed on TOE_{st} and other controls[4]. Column (4) of the table

[4] This table provides the results after we exclude the outliers. The results which include outliers (mainly big cities) are included in the Appendix.)

reports the regression coefficients for the main model specification. The results suggest that compared to the baseline year, as election approaches, the mean calibrated night light intensity increases gradually, from the coefficient going from being negative for four years before the election to being positive for one year before the election. The coefficient for four years before the election on column (4) is reported as -1.148, and for one year before the election, it is reported as 0.434. This means that for four years before the election, the mean calibrated night light intensity is 1.148 units less than the mean calibrated night light intensity for the year of election. Whereas for one year before the election, the mean calibrated night light intensity is 0.434 units greater than the election year. Both of these coefficients are statistically significant. Although the coefficients for the 2 and 3 years to the election are not statistically significant, they do show an upward pattern as the proximity to the election increases. Figure 3.13 just corroborates this fact visually. So, we see that political business cycles positively impact a proxy measure for economic development. Table C.8 in the appendix reports the results for this same regression, only the outcome variable being the change in calibrated mean night light intensity. Although we find the coefficient for one year before the election to be positive and statistically significant, any particular pattern is missing. So it is difficult to draw causal conclusions.

3.5.2 Results for analysis based on high frequency night light data and months to election

Table 3.13 and Table 3.14 shows the estimates for the different model specifications for Equation 3.2 for the median night light intensity and standardized median night light intensity as outcome variables respectively. Col 4 in Table 3.13 shows our most desired specification with the district times month fixed effects, year fixed effects, party fixed effects, demographic and economic controls and seat share difference. We find that the coefficient of MTE^{i}_{smt} increases from 3 months before the election to the election month, although only the coefficient one month before the election is statistically significant. The coefficient also decreases very sharply after the election, but the estimate is not significant. Therefore, although we find some evidence of night light intensity increasing just before the election, being positive in the election month and becoming negative in the month after election, we do not find any conclusive evidence for identifying a trend in the high-frequency night light data. Table 3.14 shows the coefficients when the median night light intensity is standardized to appreciate the magnitude of the coefficients better. The coefficients are very small, even when they are significant. Figure 3.14 and Figure 3.15 shows the coefficient estimates visually for Col 4 in Table 3.13 and Table 3.14 respectively. As seen in the figures, median night light intensity increases from two months before the election to the election month and then drops sharply afterwards, although the coefficients are really small.

To investigate the credibility of the high-frequency night light data, we aggregate the night light data and regress them on some of our demographic and economic variables, which are expected to correlate with night light positively. Table 3.15 shows the existence of a statistically significant positive correlation between median night light intensity and sex ratio, share of the population who completed secondary school and share of the population who completed high school (HS). Since these indicators of development and night light are a proxy of economic development, these significant correlations do bolster our belief in the outcome variable's credibility. The median night light intensity also negatively correlates with rural unemployment, making the outcome variable more legitimate. However, the outcome variable does not show a significant correlation with the population or with the share of the literate population, although the coefficients are positive as expected. The correlation coefficient with the share of the population who completed primary education is statistically significant and negative, but this probably makes sense since our education measures are not cumulative and completing only primary education might signify that students are dropping out before secondary school and hence can be interpreted as a measure of underdevelopment. The correlation with urban unemployment, however, is positive and significant, which is surprising. However, we can believe the high-frequency estimates are a good proxy for economic development given all the other evidence.

Having made a case for the credibility of the high-frequency night light estimates, we can argue with a greater degree of confidence that although politicians do seem to respond to the anticipation of an upcoming election just before the election, the effects are minimal and do not point towards a conclusive trend. Therefore, we conclude that there is not much of a political business cycle in the short run with respect to the night light and no evidence of very short-run gaming of electricity production.

3.6 Conclusion

In this chapter, we collect data on employment provided to labourers, road constructed, and night light intensity to investigate the effects of political business cycles on government expenditure. Although the first two data sources can be considered direct governmental expenditure, night lights can be considered a proxy for government expenditure. We find that the government's labour expenditure on mandated employment provided to labourers increases, and so does the number of people who get 100 days of employment conditional on being allotted a job as one gets closer to the election. We also find evidence of a political business cycle in new road constructed and cost sanctioned for road to be constructed, along with evidence of the political business cycle in night light intensity. Our results, therefore, provide substantial evidence of the existence of political business cycles in the micro-economic governmental expenditure in a developing country.

We also do a short analysis on a high-frequency night light data at the monthly level to investigate short-run effects of the proximity to the election. Although we do find night light intensity to have a positive coefficient just the month before the election (statistically significantly) and a positive coefficient in the month of election and a negative coefficient the month after the election, and therefore find some evidence of politicians responding to the anticipation of the election, we are unable to find any conclusive trend in the high-frequency analysis.

In further extensions of our work, we would want to look into some other outcomes such as different types of crimes committed, some health and education outcomes such as money allocated to government hospitals and schools, and investigate if the proximity to the election also influences them.

3.7 Figures and Tables

Figure 3.1: Temporal and spatial variation in total number of HH who demanded work

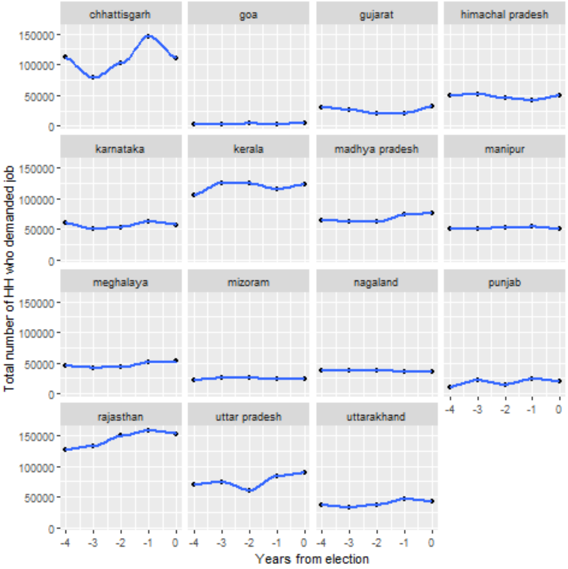
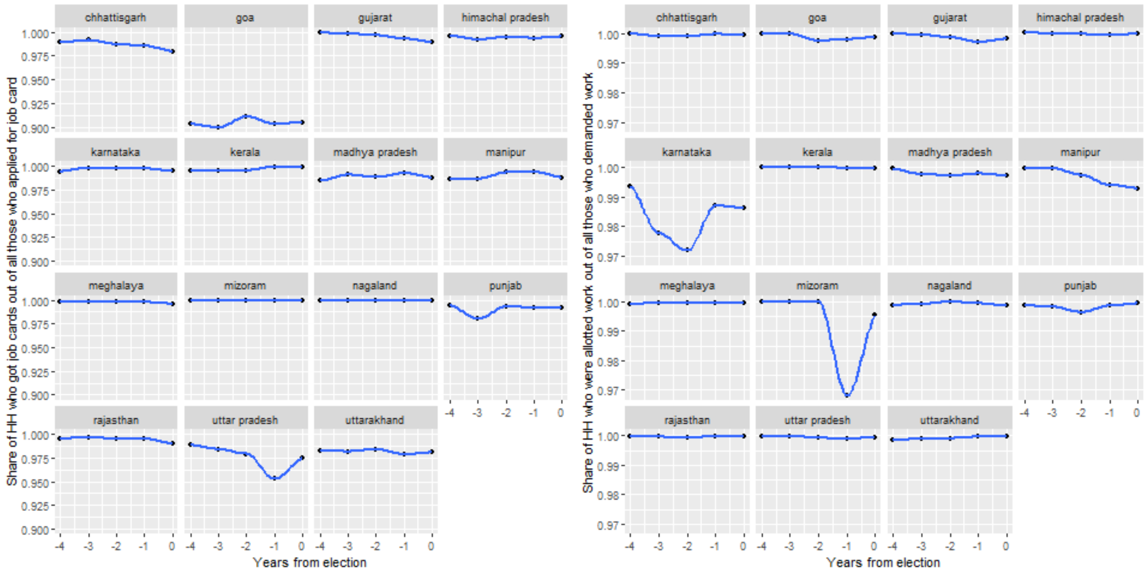


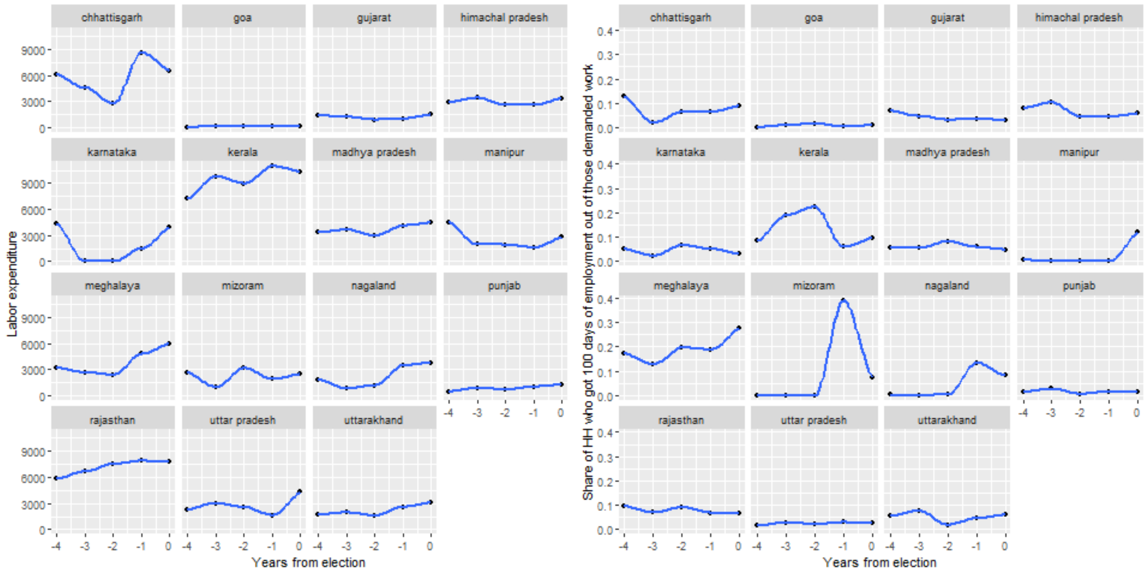
Figure 3.2: Temporal and spatial variation in share of Household we were allotted job cards and were allotted jobs



(a) Share of households who got job cards

(b) Share of households who were allotted work

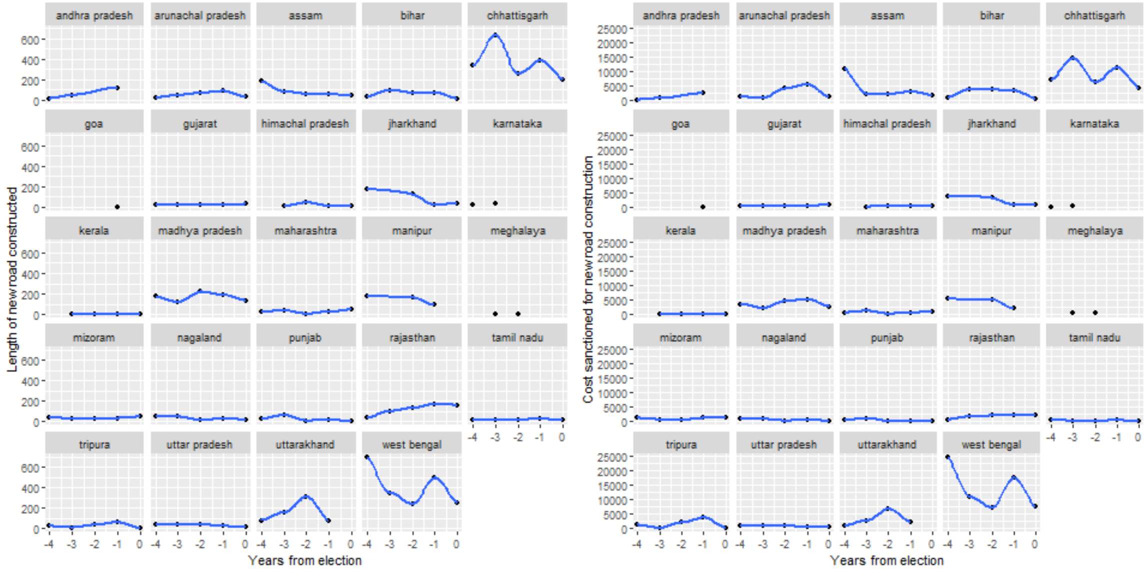
Figure 3.3: Temporal and spatial variation in labor expenditure by government and share of households who completed 100 days of employment



(a) Labor expenditure by government

(b) Share of households who completed 100 days of employment

Figure 3.4: Temporal and spatial variation in new road constructed and sanctioned cost of new road



(a) Length of new road constructed

(b) Coast sanctioned for new road

Figure 3.5: Temporal and spatial variation in mean calibrated night light intensity

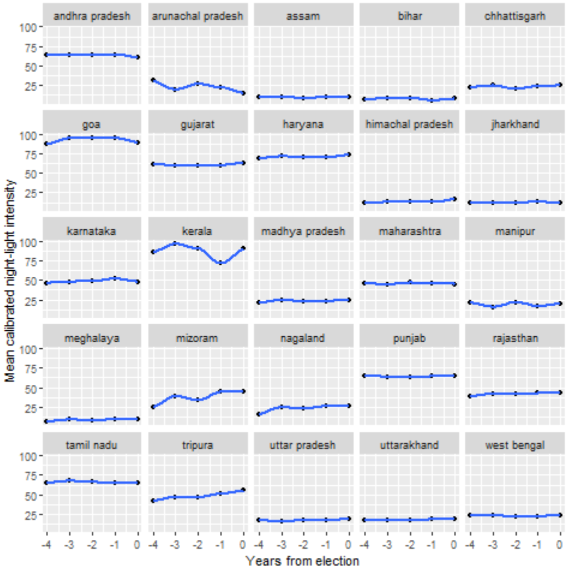


Figure 3.6

Figure 3.7: Temporal and spatial variation in high frequency median calibrated night light intensity

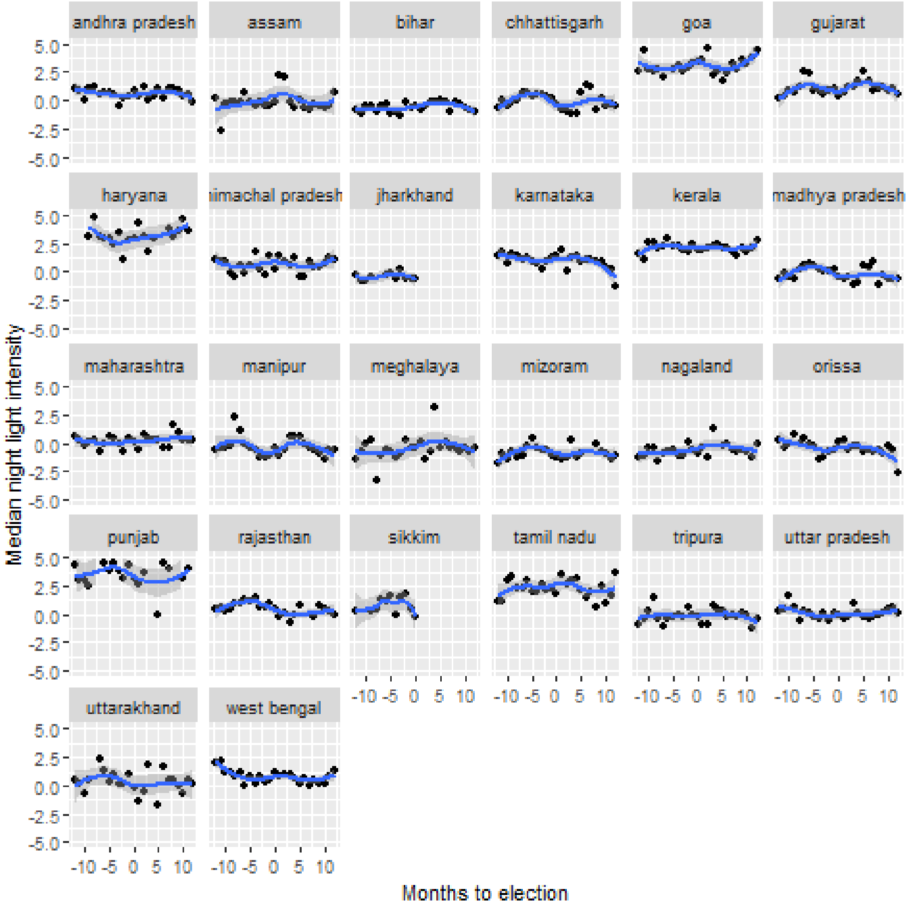


Figure 3.8: Temporal and spatial variation in standardized high frequency median calibrated night light intensity

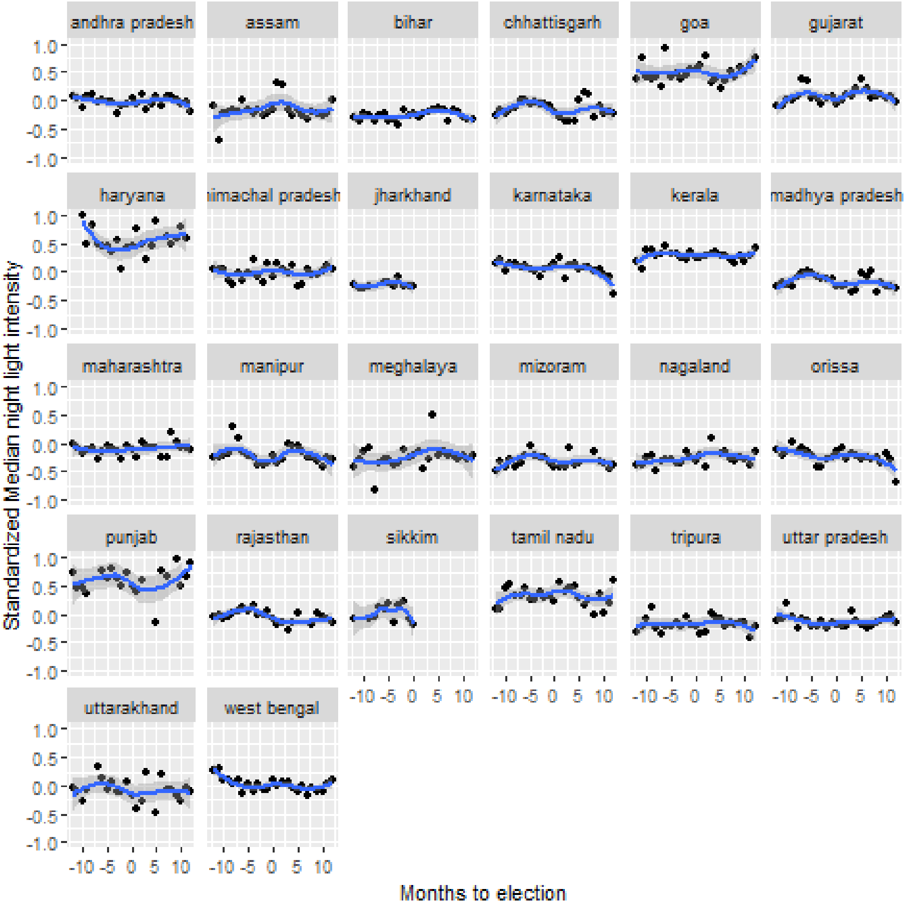
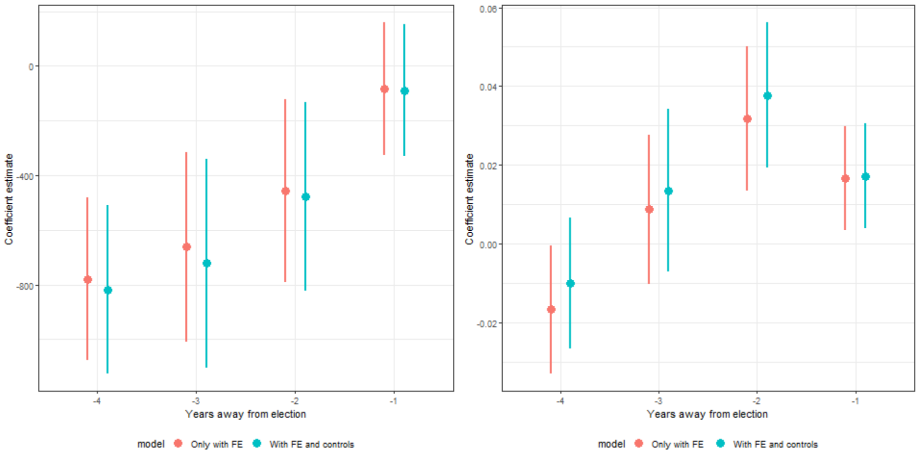


Figure 3.9: Dot Whisker Plots for coefficient estimates of labor expenditure and share of HH who completed 100 days of employment

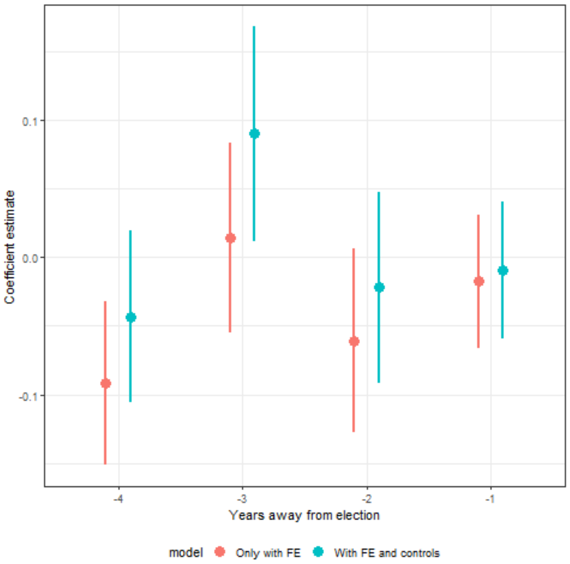


(a) Labor expenditure disbursed by government

(b) Share of HH who completed 100 days of employment

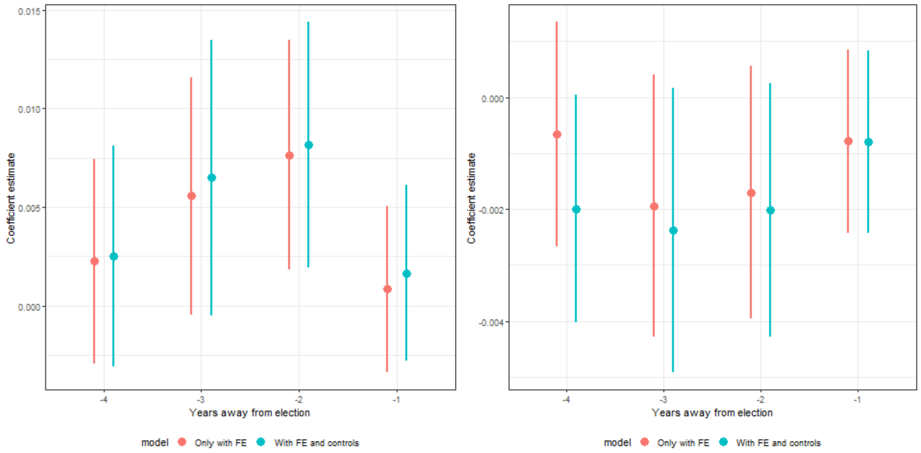
Notes: Each dot in the figure is a plot of the regression coefficient of the years to election. The orange dot represents the coefficient when the regression is ran only with the district and year fixed effects, while the blue dot represents the coefficient when the regression is ran with the fixed effects and other demographic and economic controls.

Figure 3.10: Logarithm of HH who demanded work under NREGA



Notes: Each dot in the figure is a plot of the regression coefficient of the years to election. The orange dot represents the coefficient when the regression is ran only with the district and year fixed effects, while the blue dot represents the coefficient when the regression is ran with the fixed effects and other demographic and economic controls.

Figure 3.11: Dot Whisker Plots for coefficient estimates of share of households who were issued job cards and allotted work

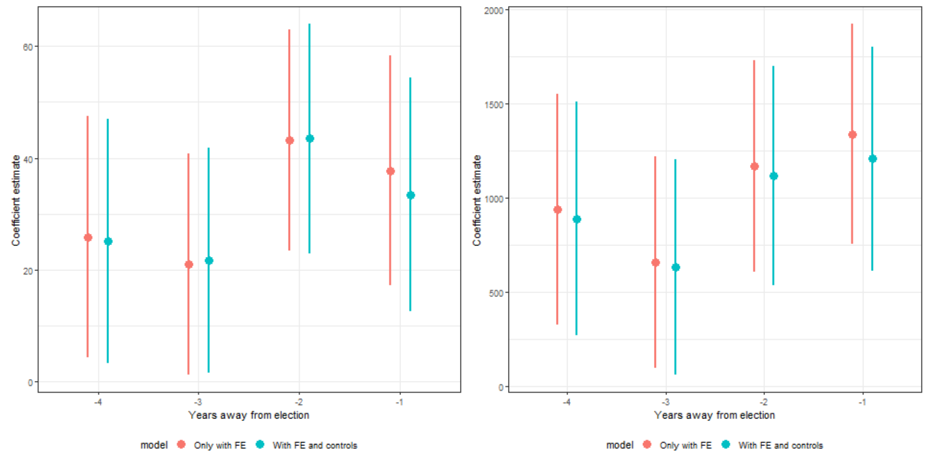


(a) Share of households who got job cards

(b) Share of households who were allotted work

Notes: Each dot in the figure is a plot of the regression coefficient of the years to election. The orange dot represents the coefficient when the regression is ran only with the district and year fixed effects, while the blue dot represents the coefficient when the regression is ran with the fixed effects and other demographic and economic controls.

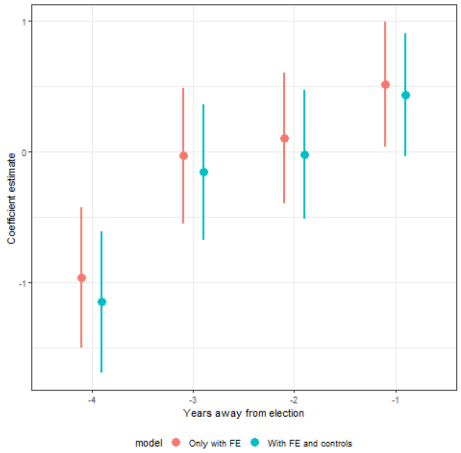
Figure 3.12: Dot Whisker Plots for coefficient estimates of new road constructed and sanctioned cost of new road



(a) Length of new road constructed (b) Cost sanctioned for new road

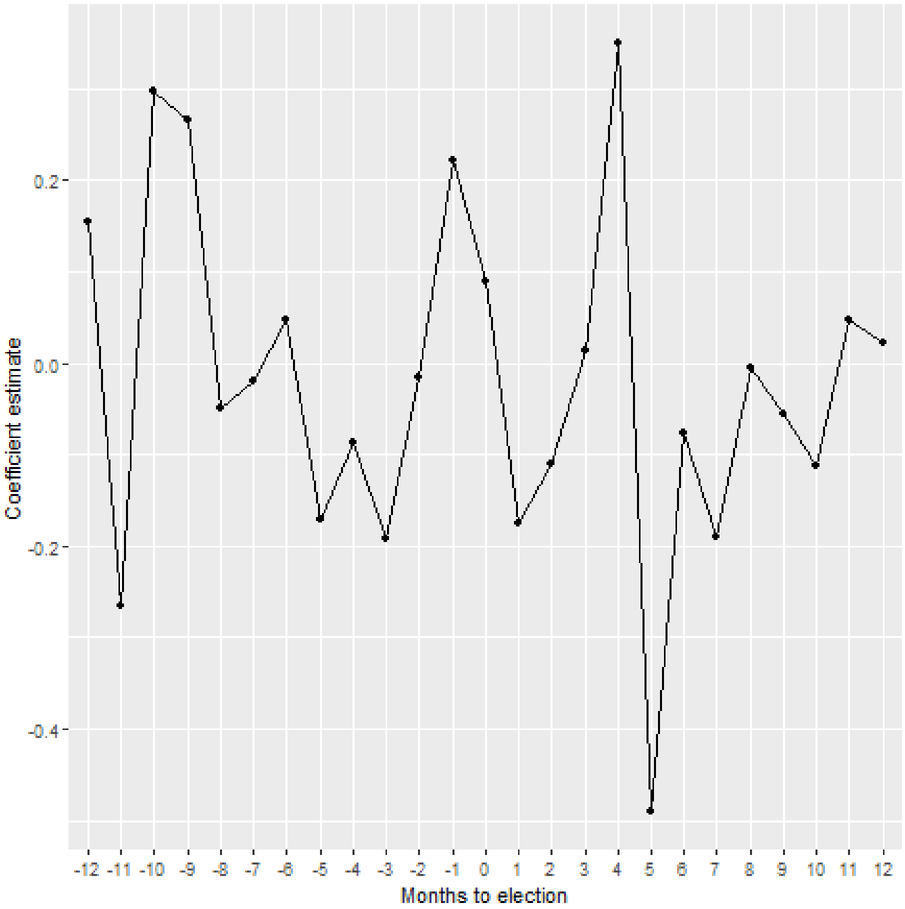
Notes: Each dot in the figure is a plot of the regression coefficient of the years to election. The orange dot represents the coefficient when the regression is ran only with the district and year fixed effects, while the blue dot represents the coefficient when the regression is ran with the fixed effects and other demographic and economic controls.

Figure 3.13: Dot Whisker Plots for coefficient estimates of mean calibrated night light intensity



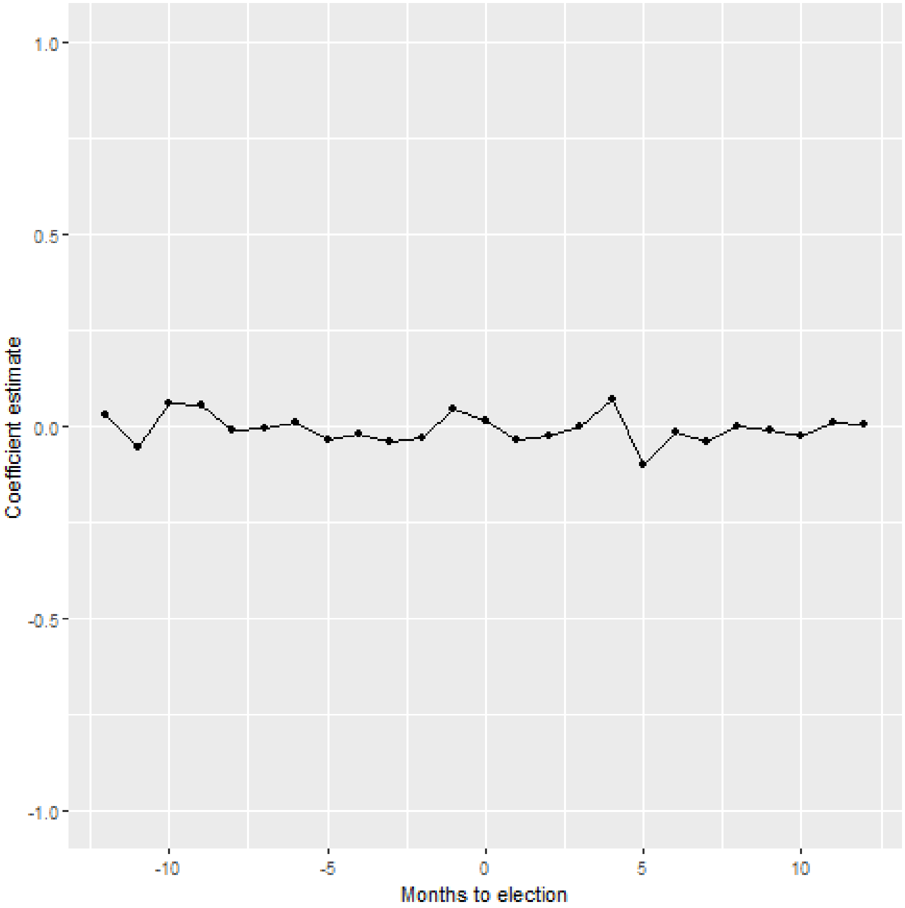
Notes: Each dot in the figure is a plot of the regression coefficient of the years to election. The orange dot represents the coefficient when the regression is ran only with the district and year fixed effects, while the blue dot represents the coefficient when the regression is ran with the fixed effects and other demographic and economic controls.

Figure 3.14: Regression coefficient plot when high frequency night light intensity is regressed on months to election.



Notes: Each dot in the figure is a plot of the regression coefficient of the months to election and demographic and economic controls, district fixed effect, (district × month fixed effect) and year fixed effects.

Figure 3.15: Regression coefficient plot when standardized high frequency night light intensity is regressed on months to election.



Notes: Each dot in the figure is a plot of the regression coefficient of the months to election and demographic and economic controls, district fixed effect, (district × month fixed effect) and year fixed effects.

Table 3.1: Summary statistics for outcome and control variables

Statistic	Mean	St. Dev.
Outcome variables		
Labor expenditure disbursed by Government (in Rs.100,000)	3,719.447	3,514.026
Total number of households who demanded work	71,612.710	56,786.270
Share of applicant HH who received job cards	0.988	0.031
Share of applicant HH who were allotted jobs	0.998	0.011
Share of HH who completed 100 days work	0.056	0.091
Length of new road constructed (in KM)	91.577	148.090
Cost sanctioned for construction of new road (in Rs.100,000)	2,195.388	4,132.280
Mean calibrated night light intensity	31.116	26.262
Control variables		
Mean number of assembly constituencies in a district	7.089	4.860
Share of constituencies won by most successful party in a district	0.655	0.213
Share of constituencies won by second most successful party in a district	0.232	0.145
Difference in share of seats between most and second most successful party in a district	0.422	0.336
Sex ratio	0.484	0.016
Share of literate population	0.546	0.141
Share of population with completed primary schooling	0.143	0.040
Share of population with completed secondary schooling	0.074	0.035
Share of population with completed higher secondary schooling	0.042	0.027
Rural unemployment per 1000 people	43.76	36.80
Urban unemployment per 1000 people	62.13	33.21
Gross State Domestic Product	42476031.73	45483560.88

Table 3.2: Summary statistics for outcome variables based on distance to election

Outcome variable	4 years to election	3 years to election	2 years to election	1 year to election	0 year to election
Labor expenditure disbursed by Government (in Rs.100,000)					
Mean(SD)	3,264.46 (2,932.69)	3,265.55 (3,350.85)	3,002.20 (3,243.36)	4,032.65 (4,021.02)	4,318.32 (3,478.17)
Median	2,444.00	2,503.00	2,029.00	2,896.00	3,546.00
Min	1	1	1	1	1
Max	18799	23107	22860	30371	23519
Total number of HH who demanded work					
Mean(SD)	65,163.13 (50,038.31)	66,050.04 (49,807.19)	64,477.36 (53,210.32)	80,345.66 (65,003.13)	74,873.09 (57,610.18)
Median	57,622.50	53,757.00	50,698.00	61,982.00	62,653.00
Min	778	1608	1080	24	49
Max	305407	285414	297910	344832	323966
Share of applicant households who received job cards					
Mean(SD)	0.99 (0.02)	0.99 (0.02)	0.99 (0.02)	0.99 (0.05)	0.99 (0.02)
Median	1.00	1.00	0.99	1.00	0.99
Min	0.813	0.821	0.838	2.854	0.801
Max	1	1	1	1	1
Share of applicant households who were allotted work					
Mean(SD)	1.00 (0.00)	1.00 (0.01)	1.00 (0.01)	1.00 (0.02)	1.00 (0.01)
Median	1.00	1.00	1.00	1.00	1.00
Min	0.987	0.950	0.943	0.647	0.926
Max	1.00	1.00	1.00	1.00	1.00
Share of applicant households who received 100 days of work					
Mean(SD)	0.05 (0.06)	0.05 (0.06)	0.06 (0.07)	0.07 (0.12)	0.05 (0.10)
Median	0.04	0.03	0.03	0.04	0.03
Min	0	0	0	0	0
Max	0.370	0.348	0.634	0.905	0.951
Length of new road constructed					
Mean(SD)	82.69 (131.10)	89.99 (144.44)	110.27 (149.95)	88.49 (176.40)	79.53 (122.34)
Median	30.10	34.57	51.62	27.45	35.80
Min	0.7	0.8	0.4	0.5	0.65
Max	916.62	1094.67	979.02	1652.36	867.5
Cost sanctioned for road construction (in Rs.100,000)					
Mean(SD)	2,010.06 (3,823.72)	2,017.61 (3,702.18)	2,495.73 (3,682.12)	2,518.70 (5,799.80)	1,752.04 (2,628.56)
Median	649.79	640.36	1,064.73	597.14	766.91
Min	4.56	13.09	0.22	12.61	8.85
Max	30689.22	24651.81	27287.76	47565.57	16344.2
Mean calibrated total night light intensity					
Mean(SD)	29.51 (25.59)	31.26 (26.47)	31.28 (26.55)	31.60 (26.31)	31.88 (26.33)
Median	20.38	22.41	22.72	23.39	23.56
Min	0.221	0	0	0	0
Max	105.0924	105.230	104.346	105.202	105.265

Table 3.3: Summary table for high frequency median night light intensity across months to election

Time to/after election	Mean(SD)	Median	Min	Max
Months to election=24	0.38 (2.61)	-0.10	-7.52	26.43
Months to election=20	0.43 (2.37)	0.05	-7.11	27.91
Months to election=16	0.60 (2.47)	0.01	-8	26.04
Months to election=12	0.71 (2.83)	0.09	-6.42	22.11
Months to election=8	0.59 (2.27)	0.20	-7.50	41.60
Months to election=4	0.55 (2.33)	0.17	-8	27.54
Months to election=1	0.63 (2.25)	0.15	-7.71	29
Months to election=0	0.46 (2.23)	0.00	-5.19	22.2
Months after election=1	0.61 (2.68)	0.17	-7.16	21
Months after election=4	0.78 (2.55)	0.21	-5.5	21.43
Months after election=8	0.73 (3.85)	0.15	-15.95	44.85
Months after election=12	0.73 (2.95)	0.20	-6.23	17.61
Months after election=16	0.60 (2.21)	0.12	-5.33	18.65
Months after election=20	0.92 (3.37)	0.16	-4.25	62.79
Months after election=24	0.69 (2.61)	0.08	-8.84	26.8

Table 3.4: Summary table for standardized high frequency median night light intensity across months to election

Time to/after election	Mean(SD)	Median	Min	Max
Months to election = 24	-0.07 (0.53)	-0.16 (-0.33, 0.12)	-1.67	5.22
Months to election = 20	-0.06 (0.48)	-0.13 (-0.30, 0.11)	-1.58	5.52
Months to election = 16	-0.02 (0.50)	-0.14 (-0.28, 0.10)	-1.76	5.14
Months to election = 12	0.00 (0.57)	-0.12 (-0.28, 0.13)	-1.44	4.34
Months to election = 8	-0.02 (0.46)	-0.10 (-0.26, 0.12)	-1.66	8.30
Months to election = 4	-0.03 (0.47)	-0.11 (-0.29, 0.09)	-1.76	5.45
Months to election = 1	-0.02 (0.46)	-0.11 (-0.27, 0.11)	-1.71	5.74
Months to election = 0	-0.05 (0.45)	-0.14 (-0.30, 0.11)	-1.19	4.36
Months after election = 1	-0.02 (0.55)	-0.11 (-0.31, 0.16)	-1.59	4.12
Months after election = 4	0.01 (0.52)	-0.10 (-0.25, 0.18)	-1.26	4.21
Months after election = 8	0.00 (0.78)	-0.11 (-0.29, 0.09)	-3.38	8.97
Months after election = 12	0.00 (0.60)	-0.10 (-0.28, 0.19)	-1.41	3.43
Months after election = 16	-0.02 (0.45)	-0.12 (-0.28, 0.12)	-1.22	3.64
Months after election = 20	0.04 (0.68)	-0.11 (-0.23, 0.11)	-1.00	12.61
Months after election = 24	-0.00 (0.53)	-0.13 (-0.27, 0.13)	-1.94	5.30

Table 3.5: The effect of proximity to election on labor expenditure made by the government in the NREGA programme

	Outcome variable:Labor expenditure disbursed by Government in Rs.100,000				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS with FE	Panel regression with FE	Panel regression with party FE	Panel regression with FE & competition	Panel regression with FE & interactions
Time to election=1 year	−83.277	−88.370	45.858	55.800	304.717
	(125.712)	(125.402)	(133.077)	(133.637)	(193.064)
Time to election=2 years	-454.965^{***}	-476.500^{***}	-323.048^{*}	-327.525^{*}	−335.705
	(153.840)	(164.324)	(170.687)	(170.231)	(215.091)
Time to election=3 years	-661.336^{***}	-721.769^{***}	-543.179^{***}	-561.660^{***}	-545.479^{**}
	(150.415)	(185.049)	(192.538)	(191.428)	(231.780)
Time to election=4 years	-779.795^{***}	-817.520^{***}	-714.188^{***}	-739.968^{***}	-770.431^{***}
	(137.806)	(143.298)	(146.279)	(145.828)	(217.792)
Difference in seat share				-381.139^{**}	−264.463
				(163.548)	(200.587)
Seat share × 1 year to election					-482.439^{*}
					(256.041)
Seat share × 2 years to election					31.690
					(277.999)
Seat share × 3 years to election					−2.800
					(254.393)
Seat share × 4 years to election					84.377
					(287.632)
Party fixed effects			✓	✓	✓
Demographic & economic controls		✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	1,642	1,642	1,642	1,642	1,642
R^2	0.853	0.855	0.858	0.858	0.859

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table 3.6: The effect of proximity to election on logarithm of households who demanded work under NREGA

	Outcome variable:Logarithm of total number of HH who demanded work				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS	Panel regression	Panel regression	Panel regression	Panel regression
	with FE	with FE	with party FE	with FE & competition	with FE & interactions
Time to election=1 year	−0.017	−0.021	−0.008	−0.009	−0.021
	(0.024)	(0.024)	(0.025)	(0.025)	(0.037)
Time to election=2 years	−0.061**	−0.042	−0.022	−0.022	−0.038
	(0.028)	(0.030)	(0.031)	(0.031)	(0.041)
Time to election=3 years	0.014	0.065**	0.088***	0.090***	0.096**
	(0.027)	(0.031)	(0.034)	(0.034)	(0.045)
Time to election=3 years	−0.092***	−0.069***	−0.046*	−0.043	−0.054
	(0.024)	(0.025)	(0.027)	(0.027)	(0.041)
Difference in seat share				0.042	0.032
				(0.037)	(0.050)
Seat share × 1 year to election					0.022
					(0.064)
Seat share × 2 years to election					0.032
					(0.060)
Seat share × 3 years to election					−0.013
					(0.065)
Seat share × 4 years to election					0.019
					(0.064)
Party fixed effects			✓	✓	✓
Demographic & economic controls		✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	1,642	1,642	1,642	1,642	1,642
R^2	0.921	0.923	0.927	0.927	0.927

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table 3.7: The effect of proximity to election on the share of job cards issued by the government in the NREGA programme

	Outcome variable:Share of job cards issued based on the number of applicants				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS	Panel regression	Panel regression	Panel regression	Panel regression
	with FE	with FE	with party FE	with FE & competition	with FE & interactions
Time to election=1 year	0.001	0.001	0.002	0.002	0.005^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
Time to election=2 years	0.008^{***}	0.007^{**}	0.008^{***}	0.008^{***}	0.007^{***}
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Time to election=3 years	0.006^{***}	0.005^{***}	0.007^{***}	0.006^{***}	0.005^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Time to election=4 years	0.002	0.002	0.003^{*}	0.003^{*}	0.002
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
Difference in seat share				-0.002^{*}	-0.001
				(0.001)	(0.002)
Seat share × 1 year to election					-0.006
					(0.004)
Seat share × 2 years to election					0.002
					(0.003)
Seat share × 3 years to election					0.003
					(0.003)
Seat share × 4 years to election					0.002
					(0.003)
Party fixed effects			✓	✓	✓
Demographic & economic controls		✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	1,642	1,642	1,642	1,642	1,642
R^2	0.421	0.430	0.432	0.432	0.433

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Notes: Standard errors are clustered at the district level.

Table 3.8: The effect of proximity to election on the share of households allotted jobs by the government in the NREGA programme

	Outcome variable:Share of households allotted jobs based on the number of applicants				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS with FE	Panel regression with FE	Panel regression with party FE	Panel regression with FE & competition	Panel regression with FE & interactions
Time to election=1 year	−0.001	−0.001	−0.001	−0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Time to election=2 years	-0.002^{**}	−0.001	-0.002^{*}	-0.002^{*}	0.0001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Time to election=3 years	-0.002^{***}	0.0002	-0.002^{*}	-0.002^{*}	−0.0001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Time to election=4 years	−0.001	0.0002	-0.002^{*}	-0.002^{*}	−0.0002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Difference in seat share				−0.002	0.0003
				(0.001)	(0.001)
Seat share × 1 year to election					−0.003
					(0.003)
Seat share × 2 years to election					-0.004^{***}
					(0.002)
Seat share × 3 years to election					-0.005^{***}
					(0.001)
Seat share × 4 years to election					-0.004^{**}
					(0.002)
Party fixed effects			✓	✓	✓
Demographic & economic controls		✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	1,642	1,642	1,642	1,642	1,642
R^2	0.252	0.301	0.356	0.357	0.360

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Notes: Standard errors are clustered at the district level.

Table 3.9: The effect of proximity to election on the share of households allotted 100 days of work by the government in the NREGA programme

	Outcome variable:Share of HH who were allotted 100 days of work out of all HH who were allotted work				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS	Panel regression	Panel regression	Panel regression	Panel regression
	with FE	with FE	with party FE	with FE & competition	with FE & interactions
Time to election=1 year	0.017**	0.018**	0.017**	0.017**	0.019
	(0.007)	(0.007)	(0.007)	(0.008)	(0.012)
Time to election=2 years	0.032***	0.030***	0.038***	0.038***	0.043***
	(0.008)	(0.010)	(0.010)	(0.010)	(0.013)
Time to election=3 years	0.009	-0.001	0.014	0.014	0.014
	(0.008)	(0.013)	(0.012)	(0.012)	(0.015)
Time to election=4 years	-0.017^{*}	-0.021^{**}	-0.010	-0.010	-0.018
	(0.009)	(0.010)	(0.010)	(0.009)	(0.012)
Difference in seat share				-0.004	-0.003
				(0.014)	(0.016)
Seat share × 1 year to election					-0.004
					(0.018)
Seat share × 2 years to election					-0.012
					(0.015)
Seat share × 3 years to election					-0.002
					(0.014)
Seat share × 4 years to election					0.015
					(0.016)
Party fixed effects			✓	✓	✓
Demographic & economic controls		✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	1,642	1,642	1,642	1,642	1,642
R^2	0.347	0.369	0.432	0.432	0.432

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Notes: Standard errors are clustered at the district level.

Table 3.10: The effect of proximity to election on the length of the road constructed by the government in the PMGSY programme

	Outcome variable:Length of new road constructed in kilometers(KM)				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS	Panel regression	Panel regression	Panel regression	Panel regression
	with FE	with FE	with party FE	with FE & competition	with FE & interactions
Time to election=1 year	23.917**	37.739***	33.321***	33.442***	11.786
	(10.579)	(12.051)	(11.802)	(11.770)	(14.043)
Time to election=2 years	30.057***	43.213***	43.092***	43.447***	28.928**
	(7.676)	(8.491)	(8.724)	(8.744)	(13.014)
Time to election=3 years	11.244	21.013**	21.650**	21.759**	15.355
	(8.994)	(9.199)	(9.115)	(9.105)	(14.677)
Time to election=4 years	7.764	25.909***	25.223**	25.166**	13.448
	(8.664)	(9.855)	(9.940)	(9.952)	(13.168)
Difference in seat share				−8.872	−37.692
				(12.456)	(27.209)
Seat share × 1 year to election					58.271
					(37.920)
Seat share × 2 years to election					37.498
					(30.761)
Seat share × 3 years to election					13.408
					(32.926)
Seat share × 4 years to election					25.540
					(32.681)
Party fixed effects			✓	✓	✓
Demographic & economic controls	✓	✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	1,598	1,598	1,598	1,598	1,598
R^2	0.137	0.026	0.063	0.063	0.067

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table 3.11: The effect of proximity to election on the cost sanctioned for new road to be constructed by the government in the PMGSY programme

	Outcome variable:Cost sanctioned for new road construction in Rs.100,000				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS	Panel regression	Panel regression	Panel regression	Panel regression
	with FE	with FE	with party FE	with FE & competition	with FE & interactions
Time to election=1 year	1,184.360***	1,336.853***	1,208.585***	1,207.276***	712.508*
	(340.010)	(358.552)	(348.197)	(347.583)	(408.549)
Time to election=2 years	787.617***	1,167.873***	1,121.056***	1,117.252***	745.285**
	(198.313)	(237.320)	(249.040)	(250.164)	(337.837)
Time to election=3 years	368.514*	655.222***	633.378**	632.210**	564.906
	(221.773)	(240.190)	(251.786)	(251.806)	(363.484)
Time to election=4 years	471.000**	939.715***	887.674***	888.279***	615.774*
	(234.003)	(236.168)	(242.206)	(242.452)	(332.411)
Difference in seat share				95.294	−539.985
				(332.161)	(702.026)
Seat share × 1 year to election					1,335.723
					(1,066.920)
Seat share × 2 years to election					949.827
					(755.155)
Seat share × 3 years to election					90.989
					(781.559)
Seat share × 4 years to election					582.767
					(817.704)
Party fixed effects			✓	✓	✓
Demographic & economic controls	✓	✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	1,598	1,598	1,598	1,598	1,598
R^2	0.115	0.027	0.068	0.068	0.072

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table 3.12: The effect of proximity to election on the mean calibrated night light intensity

	Outcome variable:Mean calibrated night-light data (Excluding outliers)				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS with FE	Panel regression with FE	Panel regression with party FE	Panel regression with FE & competition	Panel regression with FE & interactions
Time to election=1 year	-0.703^{***}	0.513^{***}	0.419^{***}	0.434^{***}	0.585^{***}
	(0.263)	(0.150)	(0.149)	(0.149)	(0.203)
Time to election=2 years	-1.398^{***}	0.103	−0.042	−0.022	0.043
	(0.310)	(0.175)	(0.181)	(0.183)	(0.240)
Time to election=3 years	-2.270^{***}	−0.029	−0.187	−0.157	−0.222
	(0.368)	(0.237)	(0.250)	(0.249)	(0.319)
Time to election=4 years	-4.290^{***}	-0.960^{***}	-1.204^{***}	-1.148^{***}	−0.537
	(0.452)	(0.268)	(0.292)	(0.298)	(0.363)
Difference in seat share				−1.414	−1.029
				(0.862)	(0.898)
Seat share × 1 year to election					−0.366
					(0.325)
Seat share × 2 years to election					−0.146
					(0.443)
Seat share × 3 years to election					0.167
					(0.509)
Seat share × 4 years to election					-1.473^{**}
					(0.656)
Seat share × 5 years to election					−0.710
					(1.957)
Party fixed effects			✓	✓	✓
Demographic & economic controls	✓	✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	7,699	7,699	7,699	7,699	7,699
R^2	0.205	0.113	0.149	0.151	0.152

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table 3.13: The effect of proximity to election on high frequency night light data

	Outcome variable:Monthly median night light intensity				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS with FE	Panel regression with FE	Panel regression with party FE	Panel regression with FE & competition	Panel regression with FE & interactions
12 months to election	0.062	0.060	0.154	0.154	0.013
	(0.047)	(0.047)	(0.110)	(0.110)	(0.178)
11 months to election	−0.079	−0.074	−0.266**	−0.266**	−0.343
	(0.055)	(0.055)	(0.126)	(0.126)	(0.231)
10 months to election	0.044	0.048	0.297***	0.297***	0.412**
	(0.056)	(0.056)	(0.104)	(0.104)	(0.187)
9 months to election	0.122***	0.127***	0.266***	0.266***	0.481***
	(0.042)	(0.041)	(0.091)	(0.091)	(0.142)
8 months to election	0.062	0.059	−0.050	−0.050	0.225
	(0.043)	(0.043)	(0.130)	(0.130)	(0.313)
7 months to election	−0.011	−0.013	−0.020	−0.020	−0.096
	(0.043)	(0.044)	(0.082)	(0.082)	(0.155)
6 months to election	0.174***	0.124**	0.048	0.048	0.321
	(0.057)	(0.057)	(0.100)	(0.100)	(0.199)
5 months to election	−0.145***	−0.180***	−0.171**	−0.171**	−0.031
	(0.041)	(0.040)	(0.068)	(0.068)	(0.120)
4 months to election	−0.136***	−0.130***	−0.087	−0.087	0.093
	(0.042)	(0.042)	(0.085)	(0.085)	(0.144)
3 months to election	−0.199***	−0.204***	−0.192**	−0.192**	−0.228
	(0.046)	(0.046)	(0.097)	(0.097)	(0.149)
2 months to election	−0.151***	−0.175***	−0.015	−0.015	−0.279*
	(0.040)	(0.039)	(0.084)	(0.084)	(0.153)
1 month to election	0.143***	0.129***	0.221***	0.221***	0.267*
	(0.035)	(0.034)	(0.075)	(0.075)	(0.144)
0 months to election	−0.097**	−0.125***	0.090	0.090	0.052
	(0.041)	(0.041)	(0.087)	(0.087)	(0.168)
1 month after election	0.070	0.067	−0.175	−0.175	−0.134
	(0.052)	(0.052)	(0.112)	(0.112)	(0.206)
2 months after election	−0.013	−0.012	−0.110	−0.110	0.052
	(0.048)	(0.048)	(0.105)	(0.105)	(0.190)
3 months after election	0.112	0.087	0.013	0.013	0.131
	(0.070)	(0.070)	(0.121)	(0.121)	(0.211)
4 months after election	0.252***	0.233***	0.350***	0.350***	0.451**
	(0.048)	(0.048)	(0.115)	(0.115)	(0.197)
5 months after election	0.040	0.020	−0.491***	−0.491***	−0.525**
	(0.055)	(0.055)	(0.113)	(0.113)	(0.212)
6 months after election	−0.125**	−0.134***	−0.076	−0.076	−0.379***
	(0.049)	(0.050)	(0.077)	(0.077)	(0.140)
7 months after election	−0.276***	−0.275***	−0.189**	−0.189**	−0.0003
	(0.046)	(0.046)	(0.079)	(0.079)	(0.155)
8 months after election	−0.021	−0.025	−0.004	−0.004	0.008
	(0.084)	(0.085)	(0.081)	(0.081)	(0.134)
9 months after election	0.062	0.076*	−0.055	−0.055	−0.237**
	(0.041)	(0.041)	(0.070)	(0.070)	(0.117)
10 months after election	−0.150***	−0.143***	−0.112	−0.112	−0.405***
	(0.041)	(0.040)	(0.080)	(0.080)	(0.125)
11 months after election	−0.042	−0.046	0.047	0.047	0.028
	(0.032)	(0.033)	(0.065)	(0.065)	(0.108)
12 months after election	0.021	0.0003	0.022	0.022	0.048
	(0.050)	(0.049)	(0.105)	(0.105)	(0.169)
Party fixed effects			✓	✓	✓
Demographic & economic controls		✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Note:					*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table 3.14: The effect of proximity to election on standardized high frequency night light data

	Outcome variable:Monthly standardized median night light intensity				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS	Panel regression	Panel regression	Panel regression	Panel regression
	with FE	with FE	with party FE	with FE & competition	with FE & interactions
time_to_election)1	0.013	0.012	0.032	0.031	0.003
	(0.009)	(0.009)	(0.022)	(0.022)	(0.036)
time_to_election)2	−0.016	−0.015	−0.053**	−0.054**	−0.070
	(0.011)	(0.011)	(0.026)	(0.026)	(0.047)
time_to_election)3	0.009	0.010	0.060***	0.060***	0.084**
	(0.011)	(0.011)	(0.021)	(0.021)	(0.038)
time_to_election)4	0.025***	0.026***	0.056***	0.054***	0.098***
	(0.008)	(0.008)	(0.019)	(0.018)	(0.029)
time_to_election)5	0.013	0.013	−0.008	−0.010	0.046
	(0.009)	(0.009)	(0.026)	(0.026)	(0.064)
time_to_election)6	−0.002	−0.002	−0.002	−0.004	−0.019
	(0.009)	(0.009)	(0.017)	(0.017)	(0.031)
time_to_election)7	0.035***	0.025**	0.010	0.010	0.065
	(0.012)	(0.012)	(0.020)	(0.020)	(0.040)
time_to_election)8	−0.029***	−0.037***	−0.034**	−0.035**	−0.006
	(0.008)	(0.008)	(0.014)	(0.014)	(0.024)
time_to_election)9	−0.028***	−0.027***	−0.018	−0.018	0.019
	(0.009)	(0.009)	(0.017)	(0.017)	(0.029)
time_to_election)10	−0.040***	−0.042***	−0.041**	−0.039**	−0.046
	(0.009)	(0.009)	(0.020)	(0.020)	(0.030)
time_to_election)11	−0.031***	−0.036***	−0.004	−0.003	−0.057*
	(0.008)	(0.008)	(0.017)	(0.017)	(0.031)
time_to_election)12	0.029***	0.026***	0.045***	0.045***	0.054*
	(0.007)	(0.007)	(0.015)	(0.015)	(0.029)
time_to_election)13	−0.020**	−0.026***	0.019	0.018	0.011
	(0.008)	(0.008)	(0.018)	(0.018)	(0.034)
time_to_election)14	0.014	0.013	−0.036	−0.036	−0.027
	(0.011)	(0.011)	(0.023)	(0.023)	(0.042)
time_to_election)15	−0.003	−0.002	−0.023	−0.022	0.011
	(0.010)	(0.010)	(0.021)	(0.021)	(0.039)
time_to_election)16	0.023	0.018	0.002	0.003	0.027
	(0.014)	(0.014)	(0.024)	(0.025)	(0.043)
time_to_election)17	0.051***	0.048***	0.071***	0.071***	0.092**
	(0.010)	(0.010)	(0.023)	(0.023)	(0.040)
time_to_election)18	0.008	0.004	−0.099***	−0.100***	−0.107**
	(0.011)	(0.011)	(0.023)	(0.023)	(0.043)
time_to_election)19	−0.025**	−0.028***	−0.017	−0.015	−0.077***
	(0.010)	(0.010)	(0.016)	(0.016)	(0.029)
time_to_election)20	−0.056***	−0.057***	−0.042***	−0.038**	−0.0001
	(0.009)	(0.009)	(0.016)	(0.016)	(0.032)
time_to_election)21	−0.004	−0.006	−0.002	−0.001	0.002
	(0.017)	(0.017)	(0.016)	(0.016)	(0.027)
time_to_election)22	0.013	0.015*	−0.014	−0.011	−0.048**
	(0.008)	(0.008)	(0.014)	(0.014)	(0.024)
time_to_election)23	−0.031***	−0.030***	−0.025	−0.023	−0.082***
	(0.008)	(0.008)	(0.016)	(0.016)	(0.025)
time_to_election)24	−0.008	−0.010	0.008	0.010	0.006
	(0.007)	(0.007)	(0.013)	(0.013)	(0.022)
time_to_election)25	0.004	−0.001	0.004	0.004	0.010
	(0.010)	(0.010)	(0.021)	(0.021)	(0.034)
Party fixed effects			✓	✓	✓
Demographic & economic controls		✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Note:					*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table 3.15: The effect of proximity to election on the high frequency median night light intensity aggregated at a yearly level

	Outcome variable:Monthly night light data aggregated to yearly data			
	(1)	(2)	(3)	(4)
	Panel regression with FE	Panel regression with party FE	Panel regression with FE & competition	Panel regression with FE & interactions
log(Total population)	0.193	0.166	0.161	0.162
	(0.154)	(0.155)	(0.155)	(0.156)
Sex ratio	11.628**	19.216***	19.077***	19.049***
	(5.852)	(5.813)	(5.836)	(5.829)
Share of literate population	1.089	0.807	0.768	0.765
	(0.849)	(0.839)	(0.833)	(0.836)
Share of population with completed primary	−7.381***	−7.699***	−7.713***	−7.704***
	(0.533)	(0.554)	(0.555)	(0.556)
Share of population with completed secondary	5.373***	4.926***	4.898***	4.894***
	(0.865)	(0.931)	(0.930)	(0.930)
Share of population with completed HS	10.959***	11.427***	11.464***	11.474***
	(0.954)	(0.991)	(0.992)	(0.991)
Unemployment rate:Rural	−0.010***	−0.009***	−0.009***	−0.009***
	(0.001)	(0.001)	(0.001)	(0.001)
Unemployment rate:Urban	0.006***	0.006***	0.005***	0.006***
	(0.001)	(0.001)	(0.001)	(0.001)
Difference in seat share			0.050	0.082
			(0.034)	(0.056)
Party fixed effects		✓	✓	✓
Demographic & economic controls	✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓
Observations	9,713	9,713	9,713	9,713
R^2	0.842	0.851	0.851	0.852

*p<0.1; **p<0.05; ***p<0.01

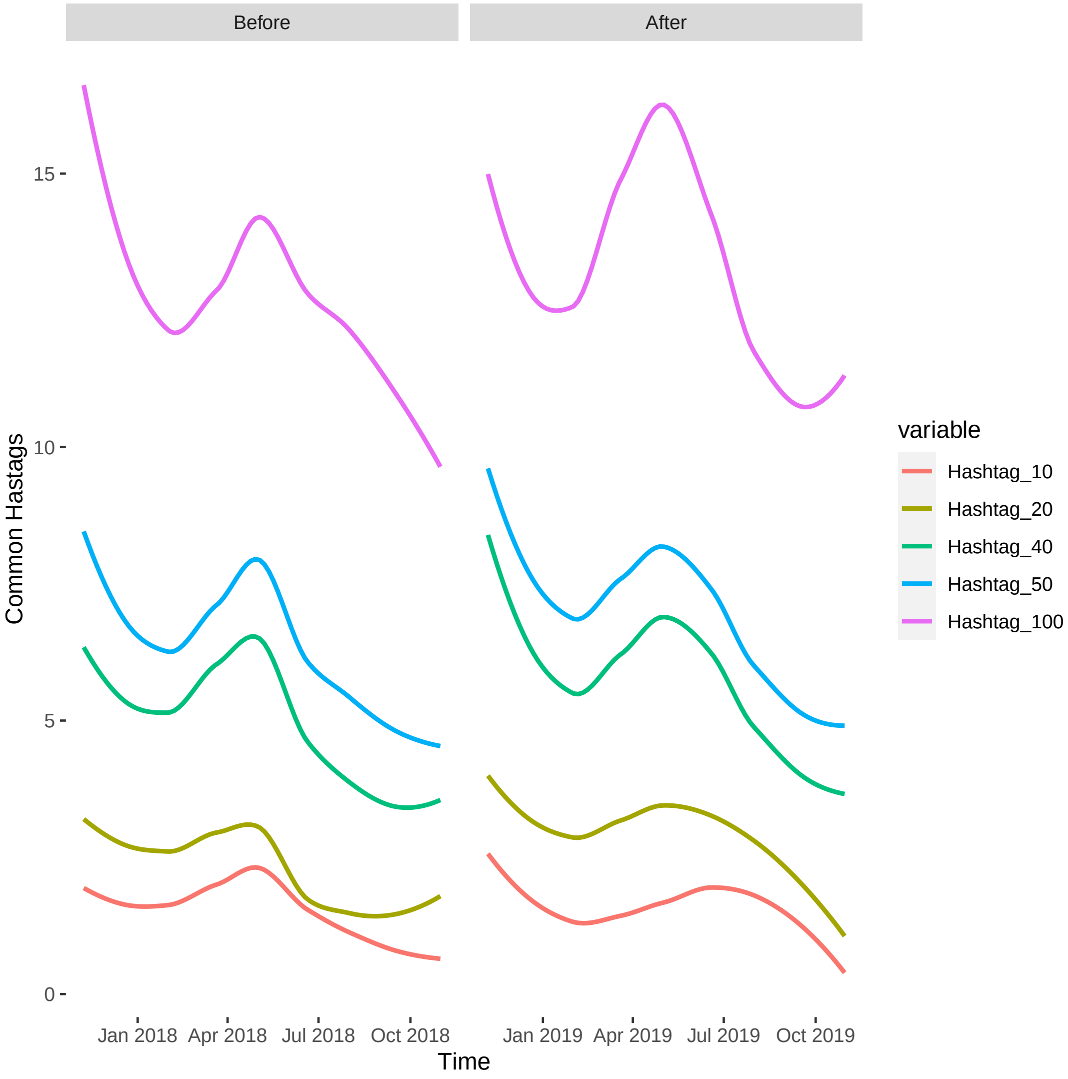
Notes: Standard errors are clustered at the district level.

Appendices

Appendix A

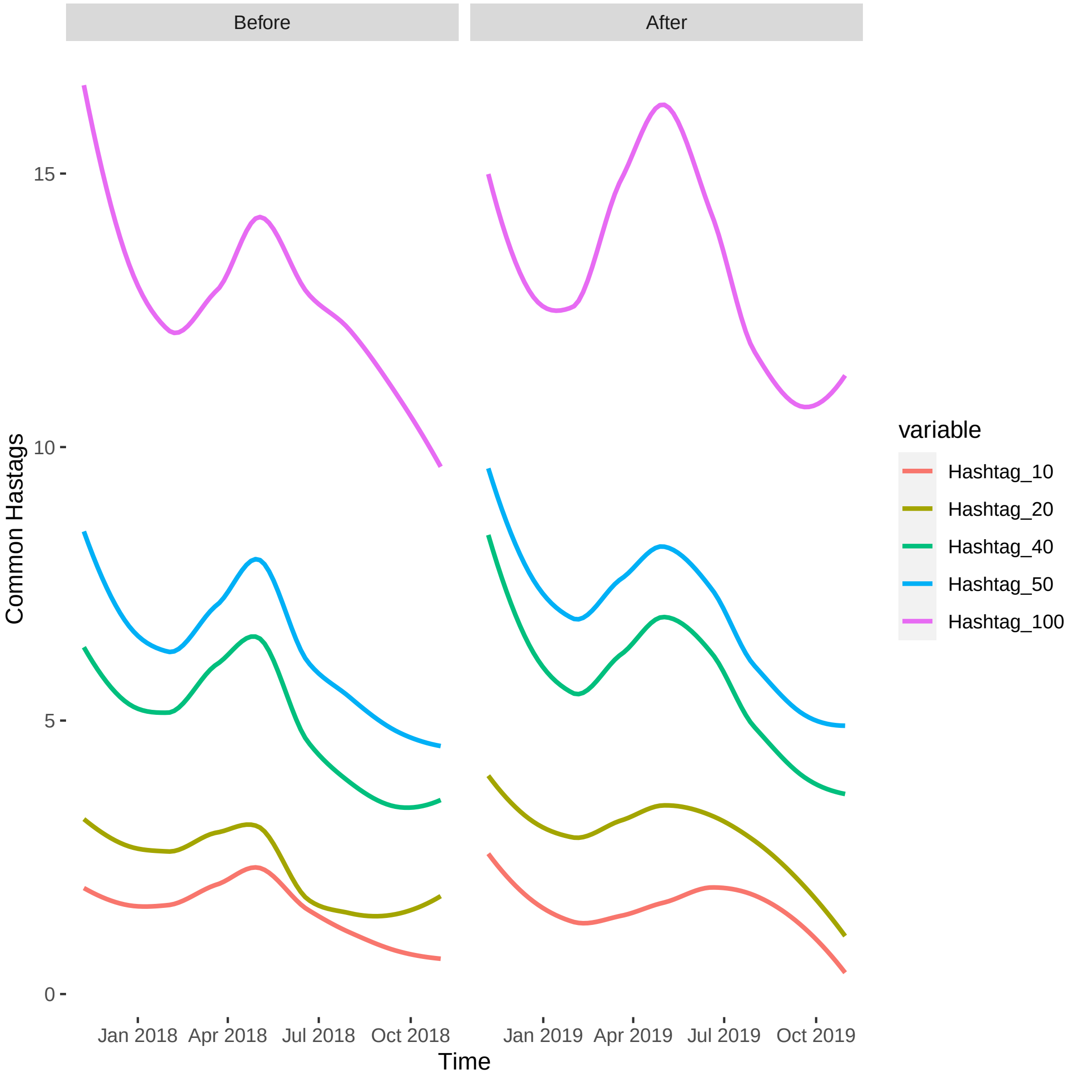
Chapter 1

Figure A.1: Similarity in hashtags for positive tweets over time



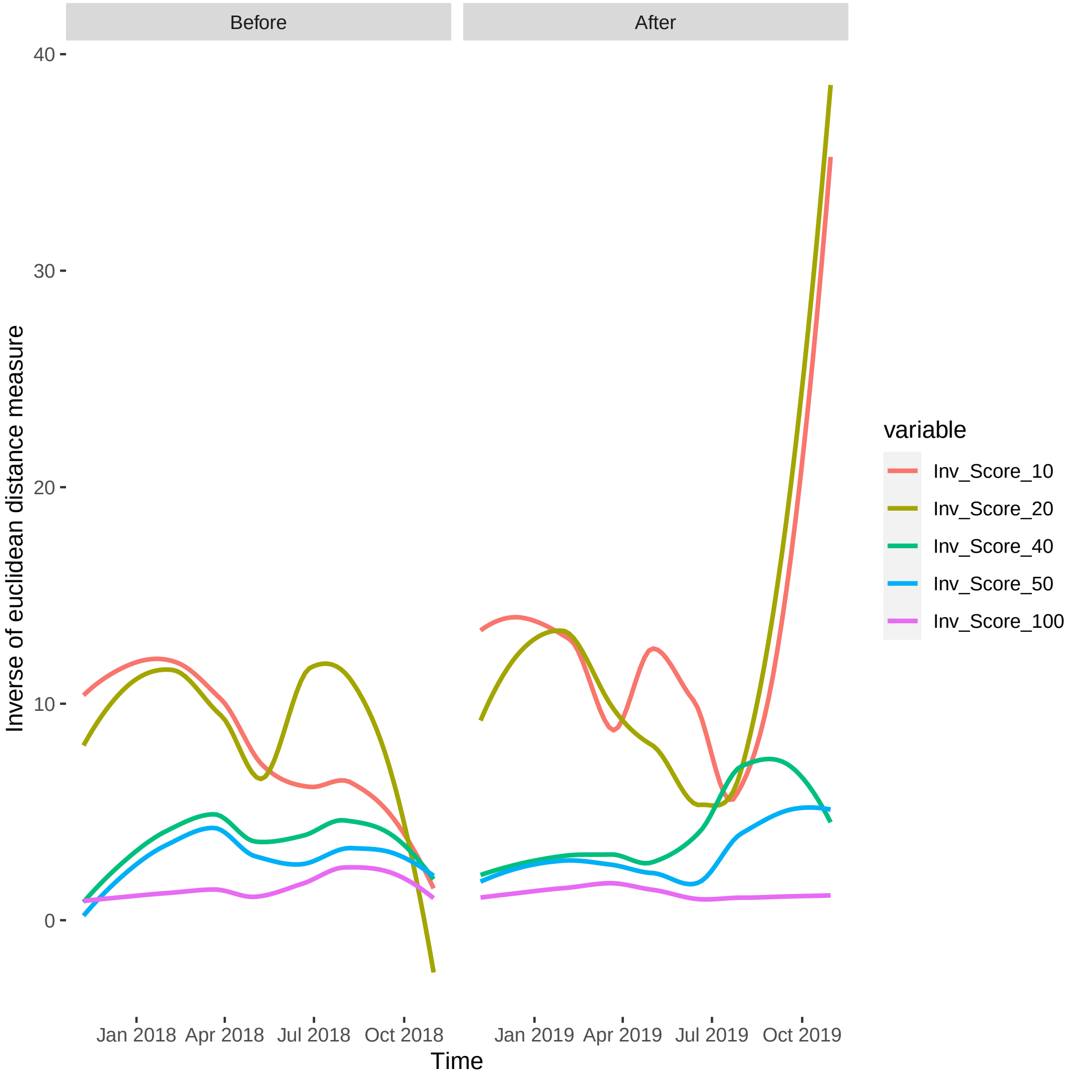
Notes: Similarity in hashtags were computed for only positive tweets, which were found by running the Sentiment Vader package on the tweets

Figure A.2: Similarity in hashtags for negative tweets over time



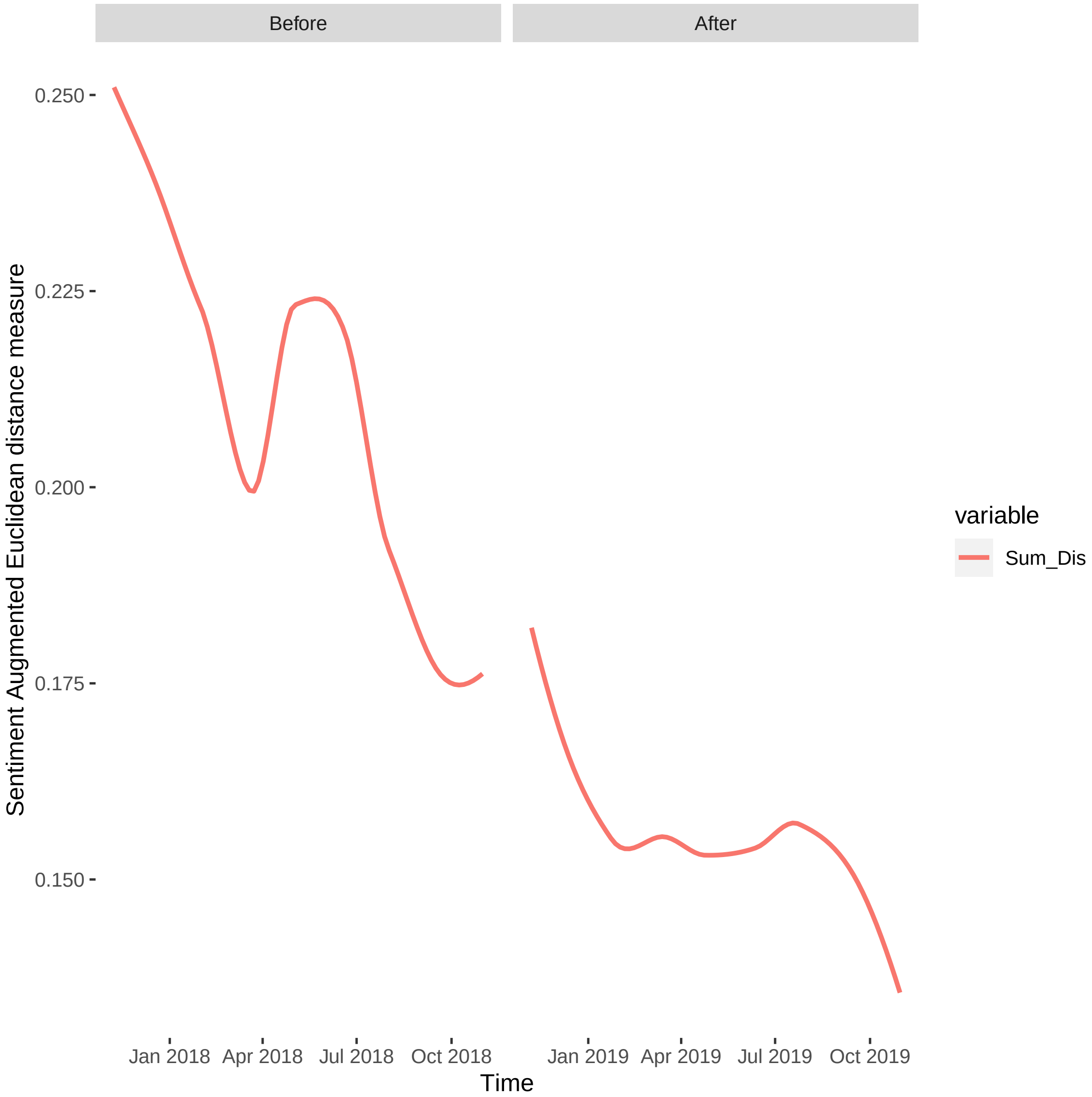
Notes: Similarity in hashtags were computed for only negative tweets, which were found by running the Sentiment Vader package on the tweets

Figure A.3: Trend of inverse of euclidean distance of hashtags interacted with sentiments



Notes: Similarity in hashtags were computed for only negative tweets, which were found by running the Sentiment Vader package on the tweets

Figure A.4: Euclidean Distance for intensity of positive, negative and neutral sentiments in a topic between Democrats and Republicans

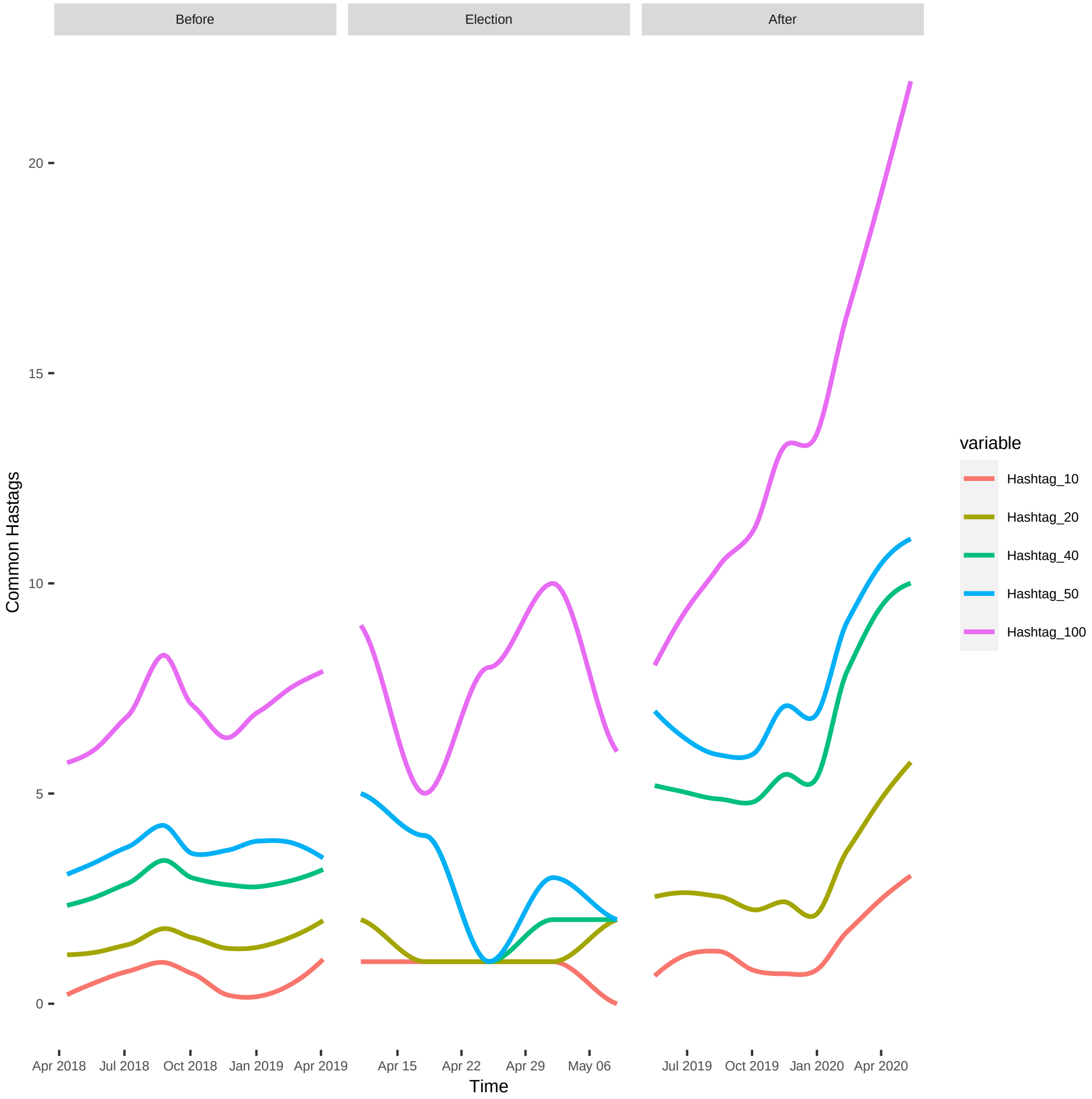


Notes: Euclidean distance between the vector of intensity of positive, negative and neutral sentiments in a topic between Democrats and Republicans. *Sum_Dis* is the euclidean distance.

Appendix B

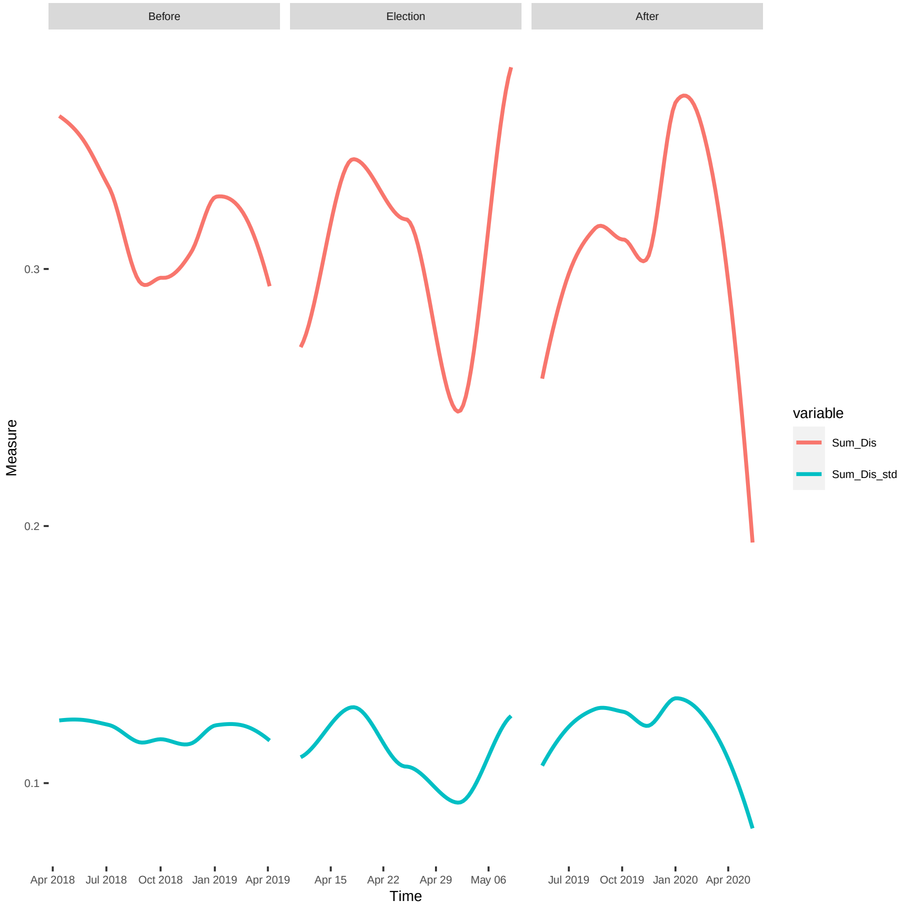
Chapter 2

Figure B.1: Similarity in hashtags over time for English and Hindi tweets



Notes: The figure plots the number of common hashtags used in the top 10, 20, 40, 50 and100 hashtags used by BJP and INC politicians in English tweets. The Before facet shows the pattern before the election, the Election period shows the pattern during the election whereas the After facet shows the pattern after election.

Figure B.2: Euclidean Distance for intensity of positive, negative and neutral sentiments in a topic between Democrats and Republicans



Notes: Sum_Dis measures the euclidean distance between the vector of intensity of positive, negative and neutral sentiments in a topic between BJP and INC politicians. Sum_Dis_{std} shows the standardized version as explained in the text. The Before facet shows what the pattern before the election, the Election period shows the pattern during the election whereas the After facet shows the pattern after election.

Appendix C

Chapter 3

Table C.1: The effect of proximity to election on change in labor expenditure made by the government in the NREGA programme

	Outcome variable:Change in labor expenditure disbursed by Government (in Rs.100,000)				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS	Panel regression	Panel regression	Panel regression	Panel regression
	with FE	with FE	with party FE	with FE & competition	with FE & interactions
Time to election=1 year	122.947	125.095	173.775	168.011	565.555***
	(127.556)	(126.170)	(138.068)	(138.412)	(198.518)
Time to election=2 years	159.313	114.619	183.082	185.678	310.289
	(195.532)	(191.599)	(196.177)	(196.137)	(263.670)
Time to election=3 years	239.391	156.997	247.294	258.010	432.989
	(214.208)	(211.288)	(212.900)	(212.922)	(294.259)
Time to election=4 years	-436.514^{**}	-471.238^{***}	-436.940^{**}	-421.992^{**}	56.143
	(180.766)	(169.320)	(169.807)	(170.603)	(246.072)
Difference in seat share				220.996	570.701***
				(140.424)	(215.126)
Seat share × 1 year to election					-768.243^{**}
					(299.122)
Seat share × 2 years to election					-230.054
					(352.208)
Seat share × 3 years to election					-303.570
					(393.003)
Seat share × 4 years to election					-920.798^{***}
					(355.273)
Party fixed effects			✓	✓	✓
Demographic & economic controls		✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	1,642	1,642	1,642	1,642	1,642
R^2	0.221	0.223	0.226	0.226	0.231

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table C.2: The effect of proximity to election on the logarithm of change in the total number of households who demanded jobs under the NREGA programm

	Outcome variable:Logarithm of change in the number of households demanding work				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS with FE	Panel regression with FE	Panel regression with party FE	Panel regression with FE & competition	Panel regression with FE & interactions
Time to election=1 year	0.062	0.057	0.027	0.018	0.045
	(0.112)	(0.110)	(0.118)	(0.117)	(0.179)
Time to election=2 years	−0.199	−0.128	−0.098	−0.087	−0.062
	(0.170)	(0.167)	(0.175)	(0.175)	(0.239)
Time to election=3 years	0.419^{**}	0.425^{**}	0.466^{**}	0.497^{**}	0.572^{**}
	(0.199)	(0.206)	(0.215)	(0.216)	(0.267)
Time to election=4 years	0.155	0.150	0.231	0.262	0.035
	(0.190)	(0.190)	(0.192)	(0.190)	(0.279)
Difference in seat share				0.312^{**}	0.293
				(0.129)	(0.188)
Seat share × 1 year to election					−0.018
					(0.263)
Seat share × 2 years to election					0.028
					(0.314)
Seat share × 3 years to election					−0.075
					(0.330)
Seat share × 4 years to election					0.592
					(0.440)
Party fixed effects			✓	✓	✓
Demographic & economic controls		✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	1,013	1,013	1,013	1,013	1,013
R^2	0.936	0.938	0.940	0.940	0.940

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table C.3: The effect of proximity to election on the change in share of job cards issued by the government in the NREGA programm

	Outcome variable:Change in the share of job cards issued based on the number of applicants				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS	Panel regression	Panel regression	Panel regression	Panel regression
	with FE	with FE	with party FE	with FE & competition	with FE & interactions
Time to election=1 year	−0.838	−0.828	−1.058	−1.063	−0.539
	(0.522)	(0.513)	(0.662)	(0.667)	(0.572)
Time to election=2 years	−0.762	−0.755	−0.963	−0.961	−0.382
	(0.573)	(0.619)	(0.754)	(0.752)	(0.745)
Time to election=3 years	−1.313	−1.333	−1.544	−1.533	−2.181
	(0.871)	(1.062)	(1.187)	(1.180)	(2.135)
Time to election=4 years	0.190	0.186	0.098	0.113	0.715
	(0.423)	(0.360)	(0.361)	(0.361)	(0.616)
Difference in seat share				0.222	0.566
				(0.258)	(0.586)
Seat share × 1 year to election					−0.989
					(0.817)
Seat share × 2 years to election					−1.142
					(1.017)
Seat share × 3 years to election					1.375
					(2.147)
Seat share × 4 years to election					−1.115
					(0.916)
Party fixed effects			✓	✓	✓
Demographic & economic controls		✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	1,642	1,642	1,642	1,642	1,642
R^2	0.163	0.163	0.164	0.164	0.166

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table C.4: The effect of proximity to election on the change in share of households allotted jobs by the government in the NREGA programme

	Outcome variable:Change in share of households allotted jobs based on the number of applicants				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS	Panel regression	Panel regression	Panel regression	Panel regression
	with FE	with FE	with party FE	with FE & competition	with FE & interactions
Time to election=1 year	−0.022	−0.022	−0.026	−0.028	0.076*
	(0.028)	(0.029)	(0.033)	(0.034)	(0.040)
Time to election=2 years	−0.041	−0.062	−0.012	−0.011	0.057
	(0.046)	(0.075)	(0.049)	(0.049)	(0.054)
Time to election=3 years	−0.018	−0.060	0.030	0.033	0.141**
	(0.063)	(0.116)	(0.065)	(0.064)	(0.067)
Time to election=4 years	−0.053	−0.073	−0.005	−0.001	0.084**
	(0.074)	(0.097)	(0.045)	(0.043)	(0.043)
Difference in seat share				0.059	0.165*
				(0.049)	(0.088)
Seat share × 1 year to election					−0.205*
					(0.106)
Seat share × 2 years to election					−0.136*
					(0.077)
Seat share × 3 years to election					−0.212***
					(0.080)
Seat share × 4 years to election					−0.161**
					(0.081)
Party fixed effects			✓	✓	✓
Demographic & economic controls		✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	1,642	1,642	1,642	1,642	1,642
R^{2}	0.158	0.161	0.223	0.223	0.228

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table C.5: The effect of proximity to election on the change in share of households allotted 100 days of work by the government in the NREGA programme

	Outcome variable:Change in share of HH who were allotted 100 days of work out of all HH who were allotted work				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS	Panel regression	Panel regression	Panel regression	Panel regression
	with FE	with FE	with party FE	with FE & competition	with FE & interactions
Time to election=1 year	13.959	11.361	18.084	18.109	68.820
	(14.483)	(12.629)	(14.130)	(14.090)	(48.874)
Time to election=2 years	25.416	14.209	25.955	25.944	46.195
	(23.859)	(14.595)	(16.803)	(16.813)	(32.875)
Time to election=3 years	35.207	25.007	31.821	31.776	51.796
	(33.063)	(22.978)	(20.550)	(20.575)	(37.171)
Time to election=4 years	45.272	39.742	29.365	29.301	46.948
	(41.309)	(35.490)	(20.044)	(20.098)	(35.353)
Difference in seta share				−0.935	35.921
				(5.830)	(30.240)
Seat share × 1 year to election					−98.933
					(68.804)
Seat share × 2 years to election					−39.041
					(33.063)
Seat share × 3 years to election					−36.062
					(33.084)
Seat share × 4 years to election					−30.728
					(29.610)
Party fixed effects			✓	✓	✓
Demographic & economic controls		✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	1,642	1,642	1,642	1,642	1,642
R^2	0.149	0.158	0.494	0.494	0.498

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table C.6: The effect of proximity to election on the change in the length of the road constructed by the government in the PMGSY programme

	Outcome variable:Change in length of new road constructed				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS with FE	Panel regression with FE	Panel regression with party FE	Panel regression with FE & competition	Panel regression with FE & interactions
Time to election=1 year	33.013^{***}	24.821^{*}	18.388	18.378	-10.700
	(12.116)	(13.858)	(13.530)	(13.518)	(19.260)
Time to election=2 years	33.786^{***}	30.863^{**}	24.571^{*}	24.512^{*}	-6.671
	(10.560)	(12.005)	(12.799)	(12.761)	(18.637)
Time to election=3 years	25.860^{***}	16.681	12.345	12.337	-6.446
	(9.797)	(11.375)	(11.682)	(11.675)	(17.718)
Time to election=4 years	26.926^{**}	21.248	18.362	18.400	-21.904
	(10.896)	(14.735)	(15.028)	(15.120)	(20.522)
Difference in seat share				2.087	-58.599^{*}
				(13.146)	(30.973)
Seat share × 1 year to election					77.183^{*}
					(40.685)
Seat share × 2 years to election					80.688^{**}
					(37.739)
Seat share × 3 years to election					45.556
					(33.914)
Seat share × 4 years to election					98.924^{**}
					(41.027)
Party fixed effects			✓	✓	✓
Demographic & economic controls	✓	✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	1,563	1,563	1,563	1,563	1,563
R^2	0.019	0.024	0.048	0.048	0.056

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table C.7: The effect of proximity to election on the change in the cost sanctioned for new road to be constructed by the government in the PMGSY programme

	Outcome variable:Change in cost sanctioned for new road construction (in Rs.100,000)				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS	Panel regression	Panel regression	Panel regression	Panel regression
	with FE	with FE	with party FE	with FE & competition	with FE & interactions
Time to election=1 year	868.005***	752.955*	490.011	488.985	46.661
	(321.894)	(385.824)	(362.837)	(362.491)	(483.416)
Time to election=2 years	653.215***	407.704	160.249	154.087	−593.493
	(242.968)	(279.778)	(305.820)	(306.888)	(418.984)
Time to election=3 years	407.593	52.084	−103.784	−104.696	−397.899
	(249.287)	(277.257)	(283.135)	(283.636)	(409.037)
Time to election=4 years	514.682**	362.565	285.044	289.028	−690.967
	(258.755)	(310.050)	(318.915)	(320.670)	(476.513)
Difference in seat share				219.891	−995.378
				(308.368)	(730.434)
Seat share × 1 year to election					1,154.186
					(1,038.350)
Seat share × 2 years to election					1,919.010**
					(876.301)
Seat share × 3 years to election					709.488
					(849.953)
Seat share × 4 years to election					2,443.643**
					(1,072.159)
Party fixed effects			✓	✓	✓
Demographic & economic controls	✓	✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	1,563	1,563	1,563	1,563	1,563
R^2	0.022	0.037	0.074	0.074	0.081

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table C.8: The effect of proximity to election on the mean calibrated night light intensity

	Outcome variable:Change in mean calibrated night-light data (Excluding outliers)				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS with FE	Panel regression with FE	Panel regression with party FE	Panel regression with FE & competition	Panel regression with FE & interactions
Time to election=1 year	0.348**	0.679***	0.674***	0.678***	0.893***
	(0.157)	(0.151)	(0.151)	(0.151)	(0.230)
Time to election=2 years	−0.451***	−0.045	−0.070	−0.064	0.138
	(0.170)	(0.162)	(0.165)	(0.166)	(0.269)
Time to election=3 years	0.467**	0.702***	0.666***	0.675***	0.117
	(0.193)	(0.181)	(0.180)	(0.179)	(0.273)
Time to election=4 years	−0.331*	0.088	0.051	0.067	0.629**
	(0.177)	(0.168)	(0.176)	(0.177)	(0.266)
Difference in seat share				−0.410**	−0.201
				(0.187)	(0.275)
Seat share × 1 year to election					−0.528
					(0.408)
Seat share × 2 years to election					−0.479
					(0.458)
Seat share × 3 years to election					1.405***
					(0.515)
Seat share × 4 years to election					−1.332***
					(0.432)
Seat share × 5 years to election					1.385*
					(0.736)
Party fixed effects			✓	✓	✓
Demographic & economic controls	✓	✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	7,699	7,699	7,699	7,699	7,699
R^{2}	0.016	0.015	0.021	0.021	0.025

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table C.9: The effect of proximity to election on the mean calibrated night light intensity including outliers

	Outcome variable:Mean calibrated night-light data (Including outliers)				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS	Panel regression	Panel regression	Panel regression	Panel regression
	with FE	with FE	with party FE	with FE & competition	with FE & interactions
Time to election=1 year	−10.551	−2.902	−2.772	−2.618	2.144
	(8.016)	(2.606)	(2.469)	(2.364)	(1.864)
Time to election=2 years	−10.215	2.493	2.816	3.042	−3.217
	(6.997)	(1.921)	(2.234)	(2.371)	(3.197)
Time to election=3 years	−10.010*	−1.189	−0.653	−0.322	−2.476
	(5.648)	(0.894)	(0.890)	(0.702)	(2.071)
Time to election=4 years	−26.976*	−4.743**	−4.009**	−3.371**	−4.251
	(14.332)	(2.354)	(1.817)	(1.396)	(2.595)
Difference in seat share				−17.382	−19.629
				(19.110)	(19.423)
Seat share × 1 year to election					−11.861
					(9.320)
Seat share × 2 years to election					15.838
					(12.723)
Seat share × 3 years to election					5.662
					(5.701)
Seat share × 4 years to election					2.123
					(4.155)
Seat share × 5 years to election					5.888
					(6.204)
Party fixed effects			✓	✓	✓
Demographic & economic controls	✓	✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	8,700	8,700	8,700	8,700	8,700
R^2	0.033	0.007	0.010	0.011	0.012

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Table C.10: The effect of proximity to election on the change in mean calibrated night light intensity including outliers

	Outcome variable:Change in mean calibrated night-light data (Including outliers)				
	(1)	(2)	(3)	(4)	(5)
	Pooled OLS	Panel regression	Panel regression	Panel regression	Panel regression
	with FE	with FE	with party FE	with FE & competition	with FE & interactions
Time to election=1 year	−9.106	−9.117	−9.075	−9.038	6.799
	(7.019)	(7.257)	(7.238)	(7.212)	(6.613)
Time to election=2 years	−3.401	−2.405	−2.302	−2.247	3.362
	(2.129)	(1.824)	(1.761)	(1.728)	(2.704)
Time to election=3 years	−0.770	0.314	0.493	0.574	2.954
	(1.199)	(0.789)	(0.697)	(0.689)	(2.277)
Time to election=4 years	−5.467	−5.186	−4.914	−4.759	1.992
	(3.375)	(3.446)	(3.292)	(3.184)	(3.575)
Difference in seat share				−4.225	11.391
				(4.587)	(8.509)
Seat share × 1 year to election					−39.723
					(31.220)
Seat share × 2 years to election					−14.255
					(10.371)
Seat share × 3 years to election					−5.952
					(6.175)
Seat share × 4 years to election					−16.753
					(13.980)
Seat share × 5 years to election					−13.810
					(11.657)
Party fixed effects			✓	✓	✓
Demographic & economic controls	✓	✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Observations	8,700	8,700	8,700	8,700	8,700
R^2	0.003	0.002	0.002	0.002	0.004

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors are clustered at the district level.

Bibliography

Adams, A. and T. McCorkindale (2013). Dialogue and transparency: A content analysis of how the 2012 presidential candidates used twitter. *Public relations review 39*(4), 357–359.

Aldrich, J. H. (1995). *Why parties?: The origin and transformation of political parties in America.* University of Chicago Press.

Alesina, A. (1988). Credibility and policy convergence in a two-party system with rational voters. *The American Economic Review 78*(4), 796–805.

Alvarez, R. M. and J. Brehm (1995). American ambivalence towards abortion policy: Development of a heteroskedastic probit model of competing values. *American Journal of Political Science*, 1055–1082.

Ambasta, P., P. V. Shankar, and M. Shah (2008). Two years of nrega: The road ahead. *Economic and Political Weekly*, 41–50.

Ash, E., D. L. Chen, and W. Lu (2017). Polarization of us circuit court judges: a machine learning approach. *Available at SSRN 2993009*.

Asher, S., T. Lunt, R. Matsuura, and P. Novosad (2020). The Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG). World Bank Economic Review (Revise and Resubmit).

Asher, S. and P. Novosad (2017). Politics and local economic growth: Evidence from india. *American Economic Journal: Applied Economics 9*(1), 229–73.

Asher, S. and P. Novosad (2020). Rural Roads and Local Economic Development. *American Economic Review).*

Auter, Z. J. and J. A. Fine (2016). Negative campaigning in the social media age: Attack advertising on facebook. *Political Behavior 38*(4), 999–1020.

Bara, J., A. Weale, and A. Bicquelet (2007). Analysing parliamentary debate with computer assistance. *Swiss Political Science Review 13*(4), 577–605.

Barberá, P., A. Casas, J. Nagler, P. J. EGAN, R. Bonneau, J. T. Jost, and J. A. Tucker (2018). Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, 1–19.

Bartels, L. M. (1986). Issue voting under uncertainty: An empirical test. *American Journal of Political Science*, 709–728.

Baskaran, T., B. Min, and Y. Uppal (2015). Election cycles and electricity provision: Evidence from a quasi-experiment with indian special elections. *Journal of Public Economics 126*, 64–73.

Berger, H. and U. Woitek (1997). Searching for political business cycles in germany. *Public Choice 91*(2), 179–197.

Besley, T. and A. Case (2003). Political institutions and policy choices: evidence from the united states. *Journal of Economic Literature 41*(1), 7–73.

Besley, T. and S. Coate (1997). An economic model of representative democracy. *The Quarterly Journal of Economics 112*(1), 85–114.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning research 3*(Jan), 993–1022.

Boyd, D. M. and N. B. Ellison (2007). Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication 13*(1), 210–230.

Boyd-Graber, J., Y. Hu, D. Mimno, et al. (2017). Applications of topic models. *Foundations and Trends® in Information Retrieval 11*(2-3), 143–296.

Bruns, A. and J. Burgess (2015). Twitter hashtags from ad hoc to calculated publics. *Hashtag publics: The power and politics of discursive networks*, 13–28.

Bruns, A., T. Highfield, and J. Burgess (2014). The arab spring and its social media audiences: English and arabic twitter users and their networks. In *Cyberactivism on the participatory web*, pp. 96–128. Routledge.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust data-driven inference in the regression-discontinuity design. *The Stata Journal 14*(4), 909–946.

Calvert, R. L. (1985). Robustness of the multidimensional voting model: Candidate motivations, uncertainty, and convergence. *American Journal of Political Science*, 69–95.

Chhibber, P. and R. Verma (2019). The rise of the second dominant party system in india: Bjp's new social coalition in 2019. *Studies in Indian Politics 7*(2), 131–148.

Clots-Figueras, I. (2012). Are Female Leaders Good for Education?Evidence from India. *American Economic Journal:Applied Economics Vol. 4*, 212–44.

Cole, S. (2009). Fixing market failures or fixing elections? agricultural credit in india. *American Economic Journal: Applied Economics 1*(1), 219–50.

Collie, M. P. and J. L. Mason (2000). The electoral connection between party and constituency reconsidered: Evidence from the us house of representatives, 1972–1994. *Continuity and change in house elections*, 211–34.

Comanor, W. S. (1976). The median voter rule and the theory of political choice. *Journal of Public Economics 5*(1-2), 169–177.

Conway, B. A., K. Kenski, and D. Wang (2013). Twitter use by presidential primary candidates during the 2012 campaign. *American Behavioral Scientist 57*(11), 1596–1610.

Conway, B. A., K. Kenski, and D. Wang (2015). The rise of twitter in the political campaign: Searching for intermedia agenda-setting effects in the presidential primary. *Journal of Computer-Mediated Communication 20*(4), 363–380.

Coughlin, P. and S. Nitzan (1981). Directional and local electoral equilibria with probabilistic voting. *Journal of Economic Theory 24*(2), 226–239.

Coughlin, P. J. (1992). *Probabilistic voting theory.* Cambridge University Press.

Dasgupta, Z. (2020). What explains india's high growth phase? investment, exports and growth during the liberalization period. Technical report, Working Paper, Azim Premji University.

Davis, L. W. (2008). The effect of driving restrictions on air quality in mexico city. *Journal of Political Economy 116*(1), 38–81.

Demszky, D., N. Garg, R. Voigt, J. Zou, M. Gentzkow, J. Shapiro, and D. Jurafsky (2019). Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *arXiv preprint arXiv:1904.01596*.

Downs, A. (1957). An economic theory of political action in a democracy. *Journal of political economy 65*(2), 135–150.

Dugoua, E., R. Kennedy, and J. Urpelainen (2018). Satellite data for the social sciences: measuring rural electrification with night-time lights. *International journal of remote sensing 39*(9), 2690–2701.

Erikson, R. S. and D. W. Romero (1990). Candidate equilibrium and the behavioral model of the vote. *American Political Science Review 84*(4), 1103–1126.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–74.

Gentzkow, M. and J. M. Shapiro (2010). What drives media slant? evidence from us daily newspapers. *Econometrica 78*(1), 35–71.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2019). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica 87*(4), 1307–1340.

Gerber, A. S. and G. A. Huber (2009). Partisanship and economic behavior: Do partisan differences in economic forecasts predict real economic behavior? *American Political Science Review*, 407–426.

Gingrich, N., D. Armey, et al. (1994). Contract with america.

Glaeser, E. L., G. A. Ponzetto, and J. M. Shapiro (2005). Strategic extremism: Why republicans and democrats divide on religious values. *The Quarterly journal of economics 120*(4), 1283–1330.

Golbeck, J., J. M. Grimes, and A. Rogers (2010). Twitter use by the us congress. *Journal of the American Society for Information Science and Technology 61*(8), 1612–1621.

Gonzalez, M. d. l. A. (2002). Do changes in democracy affect the political budget cycle? evidence from mexico. *Review of Development Economics 6*(2), 204–224.

Graham, T., M. Broersma, K. Hazelhoff, and G. Van'T Haar (2013). Between broadcasting political messages and interacting with voters: The use of twitter during the 2010 uk general election campaign. *Information, communication & society 16*(5), 692–716.

Grant, W. J., B. Moon, and J. Busby Grant (2010). Digital dialogue? australian politicians' use of the social network tool twitter. *Australian Journal of Political Science 45*(4), 579–604.

Hetherington, M. J. (2001). Resurgent mass partisanship: The role of elite polarization. *American Political Science Review 95*(3), 619–631.

Hinich, M. J. (1976). Equilibrium in spatial voting: The median voter result is an artifact.

Hotelling, H. (1990). Stability in competition. In *The Collected Economics Articles of Harold Hotelling*, pp. 50–63. Springer.

Huang, C. (2011). Facebook and twitter key to arab spring uprisings: report. In *The National*, Volume 6, pp. 2–3.

Hutto, C. J. and E. Gilbert (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Jacobi, C., W. Van Atteveldt, and K. Welbers (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism 4*(1), 89–106.

Jacobson, G. C. (2000). Party polarization in national politics: The electoral connection. In *Polarized politics: Congress and the president in a partisan era*, Volume 5, pp. 17–18.

Jaffrelot, C. (2015). The modi-centric bjp 2014 election campaign: New techniques and old tactics. *Contemporary South Asia 23*(2), 151–166.

Johnson, T. J. and D. D. Perlmutter (2010). Introduction: the facebook election. *Mass Communication and Society 13*(5), 554–559.

Klein, M. W. (1993). Timing is all: Elections and the duration of united states business cycles. Technical report, National Bureau of Economic Research.

Kumar, S. (2018). Money laundering scams in india: Its impact and government regulations to control.

Lakoff, G. (2003). Framing the issues: Uc berkeley professor george lakoff tells how conservatives used language to dominate politics. uc berkeley news, october 23, 2003.

Langa, M., S. Ndelu, Y. Edwin, and M. Vilakazi (2017). # hashtag: An analysis of the# feesmustfall movement at south african universities.

Layman, G. C., T. M. Carsey, J. C. Green, R. Herrera, and R. Cooperman (2010). Activists and conflict extension in american party politics. *American Political Science Review 104*(2), 324–346.

Ledyard, J. O. (1984). The pure theory of large two-candidate elections. *Public choice 44*(1), 7–41.

Lehne, J., J. N. Shapiro, and O. V. Eynde (2018). Building connections: Political corruption and road construction in india. *Journal of Development Economics 131*, 62–78.

Lindbeck, A. (1976). Stabilization policy in open economies with endogenous politicians. *The American Economic Review 66*(2), 1–19.

Lucas, C., R. A. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley (2015). Computer-assisted text analysis for comparative politics. *Political Analysis 23*(2), 254–277.

Mamou, J., B. Ramabhadran, and O. Siohan (2007). Vocabulary independent spoken term detection. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 615–622.

Marin, A., T. Persson, and G. E. Tabellini (1990). *Macroeconomic policy, credibility and politics*, Volume 38. Taylor & Francis.

McCallum, B. T. (1978). The political business cycle: An empirical test. *Southern economic journal*, 504–515.

McCarty, N. M., K. T. Poole, and H. Rosenthal (1997). *Income redistribution and the realignment of American politics.* Aei Pr.

McConnell, C., Y. Margalit, N. Malhotra, and M. Levendusky (2018). The economic consequences of partisanship in a polarized era. *American Journal of Political Science 62*(1), 5–18.

Michelutti, L. (2007). The vernacularization of democracy: political participation and popular politics in north india. *Journal of the Royal Anthropological Institute 13*(3), 639–656.

Monti, C., A. Rozza, G. Zappella, M. Zignani, A. Arvidsson, and E. Colleoni (2013). Modelling political disaffection from twitter data. In *Proceedings of the second international workshop on issues of sentiment discovery and opinion mining*, pp. 1–9.

Moscato, D. (2016). Media portrayals of hashtag activism: A framing analysis of canada's# idlenomore movement. *Media and Communication 4*(2), 3.

Nardi Jr, D. J. (2012). "it's only words, and words are all i have": Using latent text analysis to analyze topics in philippine supreme court decisions.

Nordhaus, W. D. (1975). The political business cycle. *The review of economic studies 42*(2), 169–190.

Ordeshook, P. C. (1986). *Game theory and political theory: An introduction.* Cambridge University Press.

Osborne, M. J. and A. Slivinski (1996). A model of political competition with citizen-candidates. *The Quarterly Journal of Economics 111*(1), 65–96.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science 54*(1), 209–228.

Rogoff, K. and A. Sibert (1988). Elections and macroeconomic policy cycles. *The review of economic studies 55*(1), 1–16.

Ryoo, J. and N. Bendle (2017). Understanding the social media strategies of us primary candidates. *Journal of Political Marketing 16*(3-4), 244–266.

Small, T. A. (2011). What the hashtag? a content analysis of canadian politics on twitter. *Information, communication & society 14*(6), 872–895.

Sokolova, M., K. Huang, S. Matwin, J. Ramisch, V. Sazonova, R. Black, C. Orwa, S. Ochieng, and N. Sambuli (2016). Topic modelling and event identification from twitter textual data. *arXiv preprint arXiv:1608.02519*.

Steyvers, M., P. Smyth, M. Rosen-Zvi, and T. Griffiths (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 306–315.

Subramanian, A. (2019). India's gdp mis-estimation: Likelihood, magnitudes, mechanisms, and implications. *CID Working Paper Series*.

Towner, T. L. and D. A. Dulio (2012). New media and political marketing in the united states: 2012 and beyond. *Journal of Political Marketing 11*(1-2), 95–119.

Vergeer, M. (2015). Twitter and political campaigning. *Sociology compass 9*(9), 745–760.

Wang, X., F. Wei, X. Liu, M. Zhou, and M. Zhang (2011). Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1031–1040.

Wittman, D. (1983). Candidate motivation: A synthesis of alternative theories. *American Political science review 77*(1), 142–157.

Xiong, Y., M. Cho, and B. Boatwright (2019). Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of twitter during the# metoo movement. *Public relations review 45*(1), 10–23.