Keep your Enemies Closer: Strategic Platform Adjustments during U.S. and French Elections

Rafael Di Tella[1] Randy Kotti[2] Caroline Le Pennec[3] Vincent Pons[4]

August 2023

Abstract

We study changes in political discourse during campaigns, using a novel dataset of candidate websites for U.S. House elections, 2002-2016, and manifestos for French parliamentary and local elections, 1958-2022. We find that candidates move to the center in ideology and rhetorical complexity between the first round (or primary) and the second round (or general election). This convergence reflects candidates' strategic adjustment to their competitor, as predicted by the median voter theorem: Using an RDD, we show that candidates converge to the platform of opponents who narrowly qualified for the last round, as opposed to those who narrowly failed to qualify.

[1] Harvard Business School and NBER; rditella@hbs.edu
[2] CREST École Polytechnique; kotti.randy@gmail.com
[3] HEC Montréal; caroline.le-pennec@hec.ca
[4] Harvard Business School, CEPR and NBER; vpons@hbs.edu

For suggestions that have improved this article, we are grateful to Daniel Auer, Gustavo Bobonis, Laurent Bouton, Pierre Boyer, Micael Castanheira, Christian Cox, Ruth Dassonneville, Gemma Dipoppa, Allan Drazen, Jean-Francois Godbout, Joanne Haddad, Nicolas de Laubier Longuet-Marx, Ro'ee Levy, Pierre-Guillaume Méon, Juan Morales, Carlo Prato, Jesse Shapiro as well as seminar and conference participants at CREST, the Université de Montréal, HEC Montréal, the London School of Economics, the University of Toronto, Wilfrid Laurier University, Sao Paulo School of Economics FGV, the 2022 American Political Science Association Meeting, the 2022 Montréal Applied Microeconomics Workshop, and the 2023 Barcelona Graduate School Summer Forum. We thank Emiliano Grossman for sharing data from the French Agenda Project.

1 Introduction

A key tenet of representative democracy is that politicians' campaign platforms and the policies they implement should follow voters' preferences. A cornerstone of modern political economy, the median voter theorem predicts that electoral incentives will generate such alignment even if candidates are only interested in their own success (Hotelling, 1929; Downs, 1957). In a two-candidate race, assuming that candidates maximize votes generates two predictions: conditional on one of the candidates' proposal, the other would like to choose one that is as close as possible to the first; and both candidates will propose the policy position preferred by the median voter in equilibrium.[1] In Palfrey (1984)'s apt summary, because candidates "follow" each other, they will end up close to the "center."

Contrasting with this view of electoral politics is parties and candidates' frequent claim that they stand for ideas. As Margaret Thatcher famously declared during her 1979 campaign: "I am not a consensus politician. I am a conviction politician." Candidates' reluctance to defend propositions they do not favor may limit convergence or prevent it altogether (Wittman, 1973, 1977). In an ideal typical citizen-candidate world, voters can only elect policies because candidates are unable to propose and implement any other than their preferred one (Osborne and Slivinski, 1996; Besley and Coate, 1997). Additional forces may induce even strategic candidates to keep their distance from their rivals and prevent full convergence to the median (Grofman, 2004), including party discipline, the threat of third candidate entry, and the possibility that voters penalize flip-flopping from one election to the next or that they abstain when the choices available to them are too similar (Adams and Merrill, 2003).

A large empirical literature has investigated the extent to which politicians' positions and policies correspond to voter preferences, generally reporting positive but small correlations (e.g., Ansolabehere, Snyder and Stewart, 2001). However, these correlations do not necessarily reflect strategic behavior on the part of politicians: they may instead be driven by the influence of media coverage and other common factors on the ideology of both politicians and voters, or by the self-selection of politicians into constituencies aligned with their views. By contrast, we provide direct evidence on candidates' strategies by studying how they change their discourse during the electoral campaign. We show that candidates tend to adjust their platform to the platform of their opponent, which leads them to converge to the center. Our results provide the first test of the mechanism underlying the median

[1] The original version of the median voter theorem considers a one-dimensional policy space, in which voters choose the platform closest to their preferred policy position. Follow-up models have identified the conditions under which the prediction of convergence to the median holds when policy preferences are multi-dimensional (Davis and Hinich, 1968; Davis, Hinich and Ordeshook, 1970; Calvert, 1985). In the separate class of probabilistic voting models, where voter preferences are uncertain, office-motivated candidates competing on a multi-dimensional policy platform target the mean voter rather than the median (Lindbeck and Weibull, 1987). Despite important differences, these models share the prediction of strategic candidate convergence to the center.

voter theorem and they shed new light on the nature of electoral competition.

Our investigation focuses on elections with two rounds in the U.S. (the primary and the general election) and in France (the first and second rounds of the general election). Because two-round elections enable to observe the same candidate in the same year, but targeting two different sets of voters, they open a window into the heart of politics. We make five distinct contributions. First, we build a novel dataset including the content of thousands of French and U.S. candidates' platforms. Second, we provide systematic evidence on changes in electoral platforms during the campaign, and show that candidates tend to converge to the center between the first and second rounds. Third, we study the key force responsible for this convergence. Exploiting quasi-experimental variation in the identity of candidates qualified for the second round, we show that candidates strategically move their platform toward the platform of opponents who qualified marginally. Fourth, we ask which candidates are most likely to adjust their platform to the platform of their competitor and find that candidates move more when they have stronger incentives to do so – for instance, when their opponent is closer to the center – and when they are more successful and experienced. Fifth, while the literature has focused on candidates' ideology and, occasionally, the topics they discuss, we consider a novel dimension of political discourse: its level of complexity. We compare the amount of convergence in ideology and complexity and show that these dimensions are complements rather than substitutes.

Data on political language have long been limited to national party manifestos collected by the Manifesto Project (Merz, Regel and Lewandowski, 2016), and to presidential candidates' speeches (Woolley and Peters, 2017). Gentzkow, Shapiro and Taddy (2019) went one step further and collected speeches made by U.S. House representatives and senators. However, these data only cover election winners in the period following the election. Instead, our analysis requires measuring the campaign platforms of all candidates, including candidates who lost the general election or did not even qualify for it and thus fall under the radar of most data collection efforts. In addition, we need to observe the same candidate both in the general election and before, at a time where official records of candidates' activity are scarce. In the U.S., we rely on candidates' campaign websites and consider all House races for the period 2002-2016. While we were able to obtain the URLs of general election websites from the Library of Congress, this resource does not cover the websites of candidates that lost their primary. To overcome this challenge, we hired a team of research assistants to systematically search for primary election websites in the Wayback Machine, an archive of the Internet including over 800 billion web pages. Using this method, we collected multiple snapshots of a total of 9,000 candidate websites, both before and, for candidates running in the general election, after the primary. In France, we use individual candidate manifestos: two-page documents written by all candidates before the first round and, if present in the runoff, before the second round, and sent by the state to all voters. We observe manifestos for 57,000 candidates in more than 8,000 local and parliamentary elections from 1958 to 2022.

The inclusion of French and U.S. elections, made possible by the availability of data on candidate-level platforms in both countries, enables us to study politicians' strategies both in a two-party and a multi-party setting. On the one hand, the French multi-party setting is characterized by frequent changes in the party system, weaker party affiliations, and fuzzier party lines, which may give candidates more flexibility and stronger incentives to adjust their platform than in the U.S. two-party setting. On the other hand, the threat of third candidate entry is more serious in the French two-round elections than in the U.S., where the one-round plurality rule creates a barrier to entry. French candidates may be reluctant to move away from their initial platform if this means losing some voters to a new competitor. While the inclusion of these two settings enriches the analysis, we interpret differences in results with caution, since they could also reflect differences in the type of data (online websites vs. printed manifestos) and in the amount of time separating the two rounds (one week in France vs. multiple months in the U.S.), among other factors.

We use text analysis to locate candidate platforms on the left-right axis before the first round (or primary election) and then again before the second round (or general election). Each word is given a score depending on how often it is used by Democrats vs. Republicans (or left- vs. right-wing candidates), and each text is then given a score depending on the words it contains. As a complementary measure of ideology, we also identify the topics covered by each candidate. We exploit data from the Manifesto Project indicating which topics are covered in each paragraph of U.S. and French party platforms in order to train a text classifier and determine which topics candidates talk about.

Beyond the ideology and content of candidates' platforms, we analyze their level of complexity, a dimension which has previously been underexplored but which is a key parameter of candidates' strategies. Indeed, politicians put a lot of effort to understand their audience and to adjust the way they communicate accordingly. Our measure of complexity summarizes three indicators: syntactic complexity, semantic complexity, and conceptual complexity. Candidates' complexity may reflect their own level of education and sophistication and it may in turn seduce some voters while antagonizing others. Speaking (or writing) in a simple way increases candidates' chances to be understood even by the least educated and sophisticated citizens, and it may signal their willingness to defend the interests of "the people." On the other hand, a more complex discourse may enable candidates to convey their views with more nuance and increase their appeal among more educated voters. We complete our effort to decipher campaign language by computing the dissimilarity between the platforms of rival candidates in an unsupervised way. This last metric accounts for dimensions of language which may not be already captured by our metrics of ideology, topics, and complexity.

Our first set of analyses investigate the extent to which candidate platforms converge toward the center between the first and second rounds. In the U.S., the same candidate competes for the votes of their party base against rivals from their party, in the primary election, before disputing centrist voters to a candidate of the opposite party, in the general election. To the extent that

Downsian forces are at play, one would expect the changes in the set of competitors and in the location of the median voter to push the candidate's platform towards the center between the two rounds. We plot the distributions of candidate ideal points on the left-right ideological axis in both rounds and verify this prediction: on average, Democrat and Republican candidates' platforms shift inward by 0.33 and 0.42 standard deviation after the primary. While political parties are organized around the ideological divide, we find that changes in candidates' level of complexity follow the same pattern. Candidates with below- and above-median levels of complexity in the primary election both converge to a more intermediate level between the primary and the general election, a result which is not explained by reversion to the mean. Furthermore, changes in complexity and ideology are complements rather than substitutes: candidates who adjust more on one dimension also tend to adjust more on the second. We also find that candidates tend to diversify the topics they cover in the general election, by decreasing the prevalence of topics that were over-represented in their primary election website and by giving more space to topics that they barely mentioned initially.

In France, the same pool of voters is called to participate in the first and second rounds. Nonetheless, like in the U.S., candidates generally target different sets of voters in the two rounds. For instance, with an average six candidates in the first round, a left-wing candidate is likely to have to compete for left-wing voters against ideologically close opponents. Once that candidate reaches the second round, electoral competition should push them toward the center if other left-wing candidates are gone and their only remaining opponent is on the center or on the right. Le Pennec (2023) documents inward shifts of the distributions of the platforms of left-wing, right-wing, and far-right candidates. We reproduce this result in our larger dataset, and extend it to complexity. Like in the U.S., candidates with simple first-round platforms shift to a more complex discourse in the second round, those starting with a more complex discourse follow the opposite trajectory, and convergence in ideology and complexity go hand-in-hand. Finally, French candidates also broaden the set of policy topics they cover in their manifestos between election rounds.

In the second part of our analysis, we explore whether the overall convergence taking place between the first and second rounds can be explained by candidates' effort to adjust their platform to their second-round opponent, thereby testing the key mechanism underlying the median voter theorem. Convergence to a more centrist platform on the ideological and complexity dimensions could plausibly result from other factors, such as learning, a weakening of party discipline (allowing moderate candidates to move back to their ideal point), or changes in the types of individuals willing to contribute time or money to the campaign. To isolate the changes in candidate platforms explained by their effort to adjust to their opponent, we exploit quasi-experimental variation in the identity of that opponent. In the U.S., this variation is provided by close primaries. We ask whether Republican candidates move their platform closer to the platform of the winner of a close Democratic primary than to the platform of their runner-up. Formally, we use a regression discontinuity design with two observations for the same Republican candidate. Our outcome measures the extent to which the change in the Republican candidate's platform between the primary and the general election closes the gap with the primary platform of the Democratic winner, on the right of the threshold, and with the primary platform of the Democratic runner-up, on the left. The running variable is the winning margin of the Democratic primary, so that observations near the threshold correspond to close primaries. The size of the jump at the threshold, if any, measures the magnitude of the adjustment by the Republican candidate to their Democratic opponent, relative to the Democratic runner-up. Importantly, while candidates may in principle choose to emulate the winner of the opposite party's primary because they infer that this platform was the most appealing to voters, our RDD shuts down this channel. Since we focus on close primary elections, similar numbers of voters are located close to the positions of the two Democratic candidates, so closing the gap with the winner demonstrates candidates' strategic intent to get closer to the opponent they are facing.

We use a similar design in the French elections. In that case, first-round races in which the candidates ranked second and third have almost exactly the same vote share provide quasi-experimental variation in the identity of the opponent that the candidate ranked first in the first round will face in the second round. We test whether the first candidate moves their platform closer to the platform of the second candidate, between the first and second rounds, than to the platform of the third candidate.

Focusing first on changes in ideology, our design is best adapted to measure strategic adjustment in the French setting. To fix ideas, consider a race in which the candidate ranked first in the first round is from a centrist party, one of the two following candidates is on the left, and the other is on the right. Focusing on races in which the second and third candidates obtained nearly exactly the same vote share in the first round, our RDD tests whether the centrist candidate moves to the left when the left-wing candidate finished a marginal second in the first round, and to the right when the right-wing candidate finished ahead. We find that this is indeed the case: overall, the change in similarity between the first-ranked candidate's ideology and that of their qualified opponent, from first to second round, is 0.46 standard-deviation larger than the change in similarity with the third-ranked candidate who failed to qualify. In the U.S., the Republican candidate will generally remain to the right of both the winner and runner-up of the Democratic primary even if they converge to their opponent. Therefore, the change in the distance between their platform and the platform of the primary winner will in general be identical to the change in the distance with the platform of the runner-up. Unsurprisingly, then, we do not find any significant differential convergence to the primary winner. While we find some evidence of convergence to the topics covered by the qualified opponent in France, we do not find any significant topic convergence in the U.S. either.

By contrast, the level of complexity of U.S. House candidates is frequently located between the complexity of the top-two candidates in the primary of the opposite party. Focusing on these middle point races, we find that candidates adjust their level of complexity to their opponent in the general election. In France, instead, we do not find any evidence of strategic adjustment on that dimension. Overall, an index summarizing convergence in ideology, complexity, topics, and unsupervised text distance shows an effect of 0.41 standard deviation in U.S. elections (significant at the 1 percent level), and 0.30 standard deviation in French elections (significant at the 5 percent level).

Our third set of results investigates which candidates are most likely to adjust their platform to the platform of the opponent they face in the final round of voting, and under which conditions. We show that candidates adjust their platform more when their incentives to do so are stronger, namely when their opponent is a greater threat. Using an alternative RDD, we find that facing an opponent who is likely to rally voters in the second round (because she is more moderate, because she is the incumbent, or because her party won more votes in the past) leads a candidate to converge more. The level of strategic adjustment is not just determined by incentives but also by politicians' type: experienced politicians, who were successful in the past (e.g., they won the previous election or received more votes in the first round of the election), adjust to their opponent the most even though they least need it.

Our paper builds on a large empirical literature studying the strategies of candidates and parties. Economists and political scientists have investigated the determinants of candidates' decision to run (instead of staying out of the race) as well as the factors facilitating dropout agreements between parties (e.g., Anagol and Fujiwara, 2016; Pons and Tricaud, 2018; Granzier, Pons and Tricaud, 2023; Dano et al., 2023); the methods candidates use to contact voters (e.g., Gerber and Green, 2000; Pons, 2018; Spenkuch and Toniatti, 2018); and their efforts to increase their campaign resources, from fundraising to the mobilization of volunteers (e.g., Bouton et al., 2022; Adena and Hager, 2022). We focus on a different but crucial aspect of candidate strategies: which platform they choose and which language they use to convey their propositions.

A core assumption of spatial competition models such as the Downs-Hotelling model (Hotelling, 1929; Downs, 1957) and the probabilistic voting model (Lindbeck and Weibull, 1987; Persson and Tabellini, 2002) is that voters are more inclined to vote for politicians that are ideologically closer to them. Consistent with this assumption, there is evidence that moderate candidates tend to win more votes than those who only appeal to a small number of extreme voters (Erikson and Wright, 1980; Canes-Wrone, Brady and Cogan, 2002). For instance, using a regression discontinuity design, Hall (2015) finds that a party gets fewer votes in the general election when an extremist won the primary. Such results imply that candidates have an incentive to adjust their discourse to voters' preferences but they do not tell us whether and to what extent candidates actually respond to this incentive.

The vast majority of the papers addressing the latter question have studied the strength of the relationship between politicians' positions, as measured through surveys or based on their roll-call votes, and voters' preferences, as proxied for instance by the presidential vote shares in a certain district.

Unsurprisingly, the prediction of full convergence has been repeatedly rejected. For instance, Poole and Rosenthal (1985) show that Democratic and Republican senators from the same state generally hold different positions even though their constituents are, by construction, identical. Other studies have attempted to estimate the weight parliamentarians assign to their constituents' preferences in their congressional votes by correlating these two variables while controlling for other factors such as party ideology, and they have generally uncovered weak positive correlations (Levitt, 1996; Ansolabehere, Snyder and Stewart, 2001). However, the magnitude of these correlations could also reflect other mechanisms, such as selection: potential candidates are drawn from the population of voters, which may reduce the gap between their preferences, and those with unusual positions may reason that they are unlikely to win and refrain from running.

Data permitting, a more direct approach to study candidates' propensity to adjust their discourse to voters' preferences is to track changes in their position over time. Using national manifestos, existing studies have provided descriptive evidence that parties adjust their policy platform to changes in public opinion (see Adams (2012) for a review) and to shifts in other parties' success. For instance, mainstream parties tend to adapt their policy agenda in response to the growing success of the radical right (Abou-Chadi, 2016; Abou-Chadi and Krause, 2020). However, studies focusing on national parties' positions are constrained by their limited number of observations. By contrast, Lee, Moretti and Butler (2004) exploit the roll call votes of individual politicians. They find that a close victory by the Democratic party increases its electoral strength (due to the incumbency advantage) but fails to shift the roll call votes of politicians elected in the district to the left, suggesting that changes in voter preferences do not shape politicians' positions. Another possible interpretation is that close victories change vote shares while leaving voters' policy preferences unaffected. Strategic politicians making that interpretation would have no reason to change their position.

To understand how politicians choose and adjust their platform, there is no reason to stop at changes across elections: a lot can be learnt from changes by the same candidate across different rounds of the same election. In the U.S., for instance, the popular press is replete with anecdotes about primary candidates winning the nomination with an extreme platform that caters to their base, only to converge to the center as they advance to the general election. However, there is surprisingly little systematic evidence about the prevalence of this behavior and the magnitude of the corresponding shifts. Acree et al. (2020) focus on three U.S. presidential candidates in 2008 and 2012 and show that these candidates moderated their language between the primary and the general election. Burden (2001) shows that the roll-call votes of representatives in the 102nd legislature were more moderate after the 1992 primary election than beforehand, but this evidence is restricted to incumbents and may be unrepresentative of other elections. By contrast, we study the behavior of both incumbents and candidates of the opposing party, and we document systematic convergence in a sample including eight distinct U.S. election cycles from 2002 to 2016. We also establish systematic convergence to the ideological center between the first and second rounds of French elections by replicating and extending Le Pennec (2023)'s results. Our large sample considerably increases the external validity of our results. Furthermore, the change in scale relative to previous work ensures that we have sufficiently many observations close to the threshold, in our RDD, enabling us to provide causal evidence on the extent to which changes in candidate platforms between rounds are driven by exogenous variation in the identity of their opponent in the second round.

Unlike previous studies, we also expand our analysis of convergence between rounds in France and the U.S. to a second dimension, complexity. The complexity of political discourse has received little attention heretofore. Exceptions include Foarta and Morelli (2021) and Ash, Morelli and Vannoni (2023), who study the costs and benefits of producing complex legislative and regulatory texts (not politicians' discourse). Closer to our paper, Spirling (2016) finds that the extension of the franchise in 19th century Britain decreased the complexity of speeches made by members of Parliament and interprets this result as an attempt to appeal to poorer and less educated voters. Building on this study spanning more than eight decades, we ask whether candidates will go all the way to adjust the complexity of their discourse within the same election.

Finally, our paper speaks to a broad set of studies which show that dramatic policy changes were brought about by the inclusion of new voters due to the expansion of the franchise to poorer households (Meltzer and Richard, 1983; Husted and Kenny, 1997) and women (Miller, 2008), the enfranchisement of minorities (Cascio and Washington, 2014), the rollout of technologies facilitating the participation of uneducated voters (Fujiwara, 2015), or the adoption of compulsory voting (Fowler, 2013). The effects on policies and on social and economic outcomes documented by these papers are generally consistent with the median voter theorem prediction but they could also be explained by alternative mechanisms, such as the entry of new candidates, a change in the party in power, or increased salience of issues mattering to the new voters. While disentangling these different mechanisms is difficult, our results suggest that politicians' efforts to compete for newly enfranchised voters and court them by moving towards their preferences may have played an important role.

The remainder of the paper proceeds as follows. Sections 2 and 3 describe our data and outcomes. Section 4 provides evidence that candidates converge to the center between election rounds. Section 5 shows that this convergence is at least partly driven by candidates' adjustments to their opponent, and Section 6 concludes.

2 Setting and Data

2.1 U.S. elections and candidate websites

U.S. elections Elections for the U.S. House of Representatives occur every two years and elect 435 representatives. A plurality of votes is enough to win the election, except in Georgia and Louisiana, where runoff general elections are held when no candidate reaches a majority of votes. We rely on the general election results made available by MIT Election Data and Science Lab (2017*a*). Although candidates from third parties run in 54% of the general elections held between 2002 and 2016, 99% of these elections were won by either a Republican or a Democrat candidate and the combined vote share of these two parties was 95% on average. In other words, these elections are largely bipartisan.

Primary elections are held in many congressional districts among candidates from the Democrat and Republican parties prior to the general election. Between 2002 and 2016, 50% of general elections were preceded by at least one competitive primary, with two candidates or more competing against each other. When both parties hold a primary in a district, these two primaries are held on the same day. In ten states, a majority of votes is required to win the primary, and a runoff takes place if no candidate reaches 50% of votes (Underhill, 2017).[2] In all other states, a plurality of votes is sufficient to win and qualify for the general election. We collected the primary election dates from the Federal Election Commission's website[3] and use the primary election results gathered by Pettigrew, Owen and Wanless (2014) for the 1956 to 2010 elections, and by Miller and Camberg (2020) for the 2012 to 2018 elections.[4]

Candidate websites Campaign websites are an ideal source of data to study political messaging. Their content is directly elaborated by the candidate and their team, they target a broad audience of voters, and they paint a comprehensive picture of a candidate's program (Druckman, Kifer and Parkin, 2010). According to campaign insiders surveyed by Druckman, Kifer and Parkin (2018) between 2008 and 2016, websites remain an important element of candidates' campaigns, despite the increased use of other forms of communication such as social media. Websites are used to present candidates' background and political positions, on top of signing up volunteers and raising funds.

We collect the content of candidate websites through the Wayback Machine, a non-profit initiative that aims to keep track of the Internet's most important content.[5] The Wayback Machine enables

[2] In North Carolina, a runoff is not required unless the second-ranked candidate calls for one, which they have the possibility to do if the first-ranked candidate gets less than 30% of votes.

[3] https://www.fec.gov/introduction-campaign-finance/election-and-voting-information/2002-us-congressional-primary-election-dates-and-candidate-filing-deadlines-ballot-access/.

[4] We note that 62 primary runoffs were missing from Pettigrew, Owen and Wanless (2014). We added results from these runoffs manually.

[5] The Wayback Machine can be accessed at https://archive.org/web/.

us to view campaign websites as they appeared online at multiple points in time throughout the campaign and to browse through them. As an example, Appendix Figure A.1 shows the main page of a website capture.

In order to find the captures of a website on the Wayback Machine, we first need to identify the website's URL. Website URLs of candidates present in a *general* election are curated and archived by the Library of Congress (LoC).[6] By the time our data analysis was completed, in spring 2023, this database ran between 2002 and 2016. More recent years had been curated but were still under embargo. We linked these URLs to the corresponding candidates in our results' database using fuzzy string matching. Since the LoC only archives URLs of general election candidates, we adopt an alternative approach to find the URLs of candidates present in a *primary* election. We use the Wayback Machine's search function, which matches keywords with URLs and titles of over 820 billion archived web pages. To collect as many primary election websites as possible, we first used brute search before turning to manual search and a series of systematic manual checks. Importantly, primary election winners' websites are referenced by the LoC database – since they qualified for the general election – but primary losers' websites are not. To avoid gathering artificially more websites from primary election winners than losers, we separated the search procedure for primary and general election candidates. Therefore, websites of primary election winners are gathered twice: once using URLs from the LoC database, and once using URLs found on the Wayback Machine's search interface.

After retrieving the URLs pointing to House of Representatives' campaign websites, we scraped all the textual content displayed on the main page and all sub-pages accessible with one click from the main page, for each time capture of each website.[7]

In total, we were able to find and verify 35,427 website captures taken between the beginning of the election year and the day of the primary election for 3,185 candidates across 2,022 competitive primary elections (an average of 11 captures per candidate, and 8 sub-pages per capture). In the remaining of the paper, we refer to these websites as *primary election websites*.[8]

We were also able to collect 156,943 website captures (an average of 27 time captures per candidate and 12 sub-pages per capture) for 5,792 candidates present in a general election, across 3,036 general elections. Specifically, 44,871 captures were taken between the beginning of the election

[6] https://www.loc.gov/collections/united-states-elections-web-archive/.

[7] For more details on both the URL search and the scraping process, see Appendix A.

[8] In districts where a competitive primary is followed by a runoff (3.2% of the districts in our sample), we followed a specific set of rules to determine the date at which we stop collecting a primary election website's captures. If both the Democrat and Republican parties hold a runoff, or if one of the two parties holds a runoff while the other party holds no primary at all, we collect all the captures taken before the date of the runoff. If one party holds a primary with a runoff while the other has a primary but no runoff, the distinction between the pre-primary and post-primary periods is more ambiguous. In these cases (accounting for only 1.1% of all districts), we stopped collecting primary elections websites' captures after the date of the primary's first round.

year and the date of the primary election, and 112,072 captures were taken between the primary election and the general election.

Appendix Table A.1 provides the number of primary and general election races for which we have collected at least one website, the total number of candidates competing in these races, and the number of candidates for which a website is available.

2.2 French elections and candidate manifestos

Electoral rule We combine our U.S. dataset with data from French local and parliamentary elections held between 1958 and 2022. Parliamentary elections are held every five years to elect the 577 representatives seating in the *Assemblée Nationale*, the lower house of the French Parliament. Local elections are held every six years to elect the representatives seating in departmental councils, which hold the legislative power in each of the 101 *départements* – the territorial and administrative units in charge of education, transportation, and social assistance, among other prerogatives.

Parliamentary and local elections follow the same electoral rule: each constituency (*circonscription* for the former election type, *canton* for the latter) elects one representative under a two-round plurality rule. A candidate needs to obtain the absolute majority of the votes cast in their constituency to win in the first round, and these votes need to account for at least 25% of all registered voters. If no candidate is elected in the first round, the two candidates who received the most votes and any other candidate who obtained the votes of at least 12.5% of the registered voters qualify for the second round. The runoff takes place among all qualified candidates who choose to stay in the race, and the candidate who receives a plurality of votes gets elected.[9]

While this electoral rule has remained relatively stable over the period that we study, a few historical changes must be noted. First, the first round vote share required to qualify for the second round has changed. In parliamentary elections, it went from 5% of the expressed votes in 1958 to 10% of the registered voters in 1966 and to the current threshold of 12.5% of the registered voters in 1975. In local elections, it went from 10% of the registered voters to 12.5% in 2010. Second, the 1986 parliamentary election used a proportional list system at the *département* level instead of the two-round plurality rule in single-member constituencies described above. We exclude this election year from our sample. Third, local elections were substantially transformed through a large reform prior to the 2015 elections. Until 2015, each *canton* elected one council member for a six-year mandate and, every three years, half of the constituencies voted to renew their representatives. Since 2015, local elections have been taking place every six years in the entire country, and they elect tickets composed of a male and a female candidate and campaigning on a common platform.

[9] Over our sample period, a runoff was held in 86% of parliamentary elections and 69% of local elections.

Electoral results at the candidate level for each of these elections, along with candidate characteristics such as their gender, their incumbency status, and whether they have run in the past or not, were obtained from Granzier, Pons and Tricaud (2023) and Dano et al. (2023).

Party system Unlike the two-party U.S. setting, the French political system is multipartisan. French politics have been historically dominated by a left-wing block organized around the Socialist party and a right-wing block organized around the conservative Gaullist party (currently called *Les Républicains*), but other parties have also been present on the ballot in many constituencies throughout our sample period. Important examples include the Communist party, the Green party, the centrist party MODEM, and the far-right party RN. Beyond these large national parties, elections often feature candidates affiliated with smaller issue-specific parties. Candidates may also run as independents, without the endorsement of any national party.[10]

Following Granzier, Pons and Tricaud (2023), we allocate candidates to six political orientations (far-left, left, center, right, far-right, and non-classified) based on political labels assigned to each candidate by the Ministry of the Interior – the official publisher of the election results. This classification applies to independent candidates as well, since candidates who are not affiliated with any of the main parties might nonetheless have a clear political orientation, indicated by labels such as "diverse left" or "diverse right."

Candidate manifestos During the official campaign period preceding both local and parliamentary elections, candidates are invited to issue a campaign manifesto (or *profession de foi*), in which they may advertise their policy platform as well as their personal attributes. These two-page documents are mailed to all registered voters by the state a few days before the first round.[11] Importantly, candidates who qualify for the runoff are invited to issue a new manifesto prior to the second round, that is also sent to voters. Additional details are provided in Appendix A and an example of a candidate manifesto is shown in Appendix Figure A.2.

Our dataset of candidate manifestos builds on a corpus assembled by Le Pennec (2023) for the parliamentary elections held between 1958 and 1993. Le Pennec (2023)'s data only cover a single year post 1993: the 2017 parliamentary elections.[12]

We first complete these data with the manifestos for the 1997 parliamentary elections, which were collected by Cagé, Le Pennec and Mougin (2023).

[10] Independent candidates account for 31% of all candidates running in our sample period and for 18% of the candidates who are ranked first, second, or third in the first round.

[11] The expenditures associated with the preparation and printing of manifestos are fully reimbursed by the state, provided that the candidate obtains at least 5% of the votes in the first round of the election.

[12] The 1958-1993 manifestos were digitized by the Archelec project (Gaultier-Voituriez, 2016), and the 2017 manifestos were published online (at the candidate's discretion) on the Ministry of the Interior's website.

Second, we scraped the manifestos published online for the 2021 local election and the 2022 parliamentary election.[13]

Third, we made a big effort to find manifestos issued for the intermediate parliamentary elections, held in 2002, 2007, and 2012, and for the local elections held between 1979 and 2015. We visited departmental archives, municipal archives, and town halls throughout *Ile de France*, France's most populated region (which includes Paris and its surroundings) and were able to find and digitize a subset of the manifestos for these elections.[14] We applied optical character recognition to convert their content into machine-readable text.

In total, our dataset contains first-round manifestos issued by 46,607 candidates across 8,156 races, and second-round manifestos issued by 10,310 runoff candidates across 5,209 races. Appendix Table A.2 indicates the number of first- and second-round races for which we have at least one manifesto, for each local and parliamentary election, the total number of candidates competing in these races, and the number of candidates for which a manifesto is available.

3 The Multiple Dimensions of Political Discourse

We consider three dimensions of political discourse: ideology, the complexity of candidates' language, and the policy topics they talk about. We use text analysis to construct measures associated with each of these dimensions. We also construct an agnostic measure of text similarity between any two documents.

3.1 Ideological score

Our first measure captures the ideology of candidates' discourse. We use a supervised approach to scale candidate websites and manifestos from left to right. Intuitively, we use candidates' party affiliation to identify all words associated with each ideological side and we then estimate candidates' ideology based on their choice of words. Within each country, we treat each election year separately in order to account for changes in the ideological leaning of words over time. Our approach builds on the Wordscores method introduced by Laver, Benoit and Garry (2003).

The ideological score of document j is defined as:

$$S_j = \sum_w p_{wj} \cdot s_w,$$

[13] These manifestos were available at https://programme-candidats.interieur.gouv.fr/.

[14] See Appendix A for more details on this data collection effort.

where s_w is the score of word w, as defined below, and $p_{wj} = \frac{c_{wj}}{m_j}$ is the frequency of word w in document j, with c_{wj} the number of occurrences of word w in document j and m_j the total number of words in document j.

In France, a document is either a first-round or second-round manifesto. In the U.S., we concatenate all the different captures of a campaign website prior to the primary election date to have one primary election data point per candidate. Similarly, we concatenate all the captures of a website between the day of the primary election and the day of the general election to have one general election data point per candidate. Therefore, in both countries, we have at most two data points (and two scores) per candidate: one for each election round. In the remainder of the paper, we use the term "first round" to refer both to primary elections in the U.S. and to first election rounds in France, and the term "second round" to refer both to general elections in the U.S. and to runoffs in France.

We construct the word scores s_w in the following way. In the U.S., we use the content of websites published by Democratic candidates (labeled as "left") on one side, and Republican candidates (labeled as "right") on the other, excluding websites from independent or third-party candidates. In France, we aggregate the content of manifestos issued by candidates labeled as left-wing or right-wing, excluding manifestos from centrist and non-classified candidates.[15] We only use the primary election websites and first-round manifestos as reference texts.[16] Word scores are then defined as:

$$s_w = \frac{p_w^R}{p_w^L + p_w^R} - \frac{p_w^L}{p_w^L + p_w^R},$$

where $p_w^I = \frac{1}{|I|} \sum_{j \in I} p_{wj}$ is the average frequency of word w among documents from ideological side I (with $I = L, R$). The score of each word ranges from -1, when it is only used by left-wing candidates and never by right-wing ones, to 1, when it is only used by right-wing candidates and never by left-wing ones. We exclude words that are used too infrequently in order to limit the influence of rare words that appear extreme because they are, by chance, used only by one or few candidates from the same ideological side, even though they do not carry any partisan meaning. We also exclude the most commonly used words, which do not differentiate partisan sides, to improve computational efficiency.[17]

Intuitively, a website or a manifesto with a negative (positive) score contains mostly words that are primarily used by left (right)-wing candidates, while a document with a partisan score close

[15] Left-wing candidates include candidates from parties such as the Communist Party or the Socialist Party, and candidates labeled as "Other Left." Right-wing candidates include candidates from parties such as the National Front or Rally for the Republic, and candidates labeled as "Other Right."

[16] If candidates start using words from the opposite ideological side in the second round, using second-round documents as reference texts would result in more neutral word scores, and candidates who use these words in the first round would receive a more neutral ideological score as well. This could lead us to underestimate candidates' changes in ideological scores between rounds as well as their adjustments to the opponent.

[17] Specifically, we exclude words that are used by fewer than 0.5% and more than 80% of all left-wing and right-wing candidates.

to zero either contains words that are just as common among left-wing and right-wing candidates of words that are primarily used by the left and words primarily used by the right.[18] Appendix Tables B.1 and B.2 show the twenty words with the highest (most right-leaning) and lowest (most left-leaning) score for the U.S. and French elections in the sample. In both countries, we find words referring to economic policy and redistribution (e.g., "wealthiest," "richest," "equality," "trades") and the environment (e.g., "renewable," "solar," "pollution") on the left. In the U.S., many left-wing words also refer to minority rights (e.g., "minority," "discrimination," "gay") while many right-wing words refer to religion (e.g., "sanctity," "pray," "bible") and abortion (e.g., "unborn," "abortions"). In France, many right-wing words refer to immigration policy (e.g., "stay," "clandestine") and crime (e.g., "military police," "brigade," "terrorism").

We also compare the ideological score of U.S. election winners, based on their general election website, with their DW-Nominate score once in office, obtained from Boche et al. (2018). DW-Nominate scores are standard measures of legislators' ideal points based on their roll-call votes (Nokken and Poole, 2004). When pooling all election winners together, the correlation between the two metrics is 0.50. Within the same party, this correlation is 0.13 for Democrats and 0.16 for Republicans. Similarly, we compare the website's ideological score of incumbents running for reelection with their DW-Nominate score estimated over the term preceding the election. The correlation is 0.47 overall, 0.21 for Democrats, and 0.22 for Republicans. All correlations are significant at the 1% level.[19] Unfortunately, we cannot conduct the same exercise in France, due to the limited availability of roll-call data.

3.2 Complexity score

Our second measure captures the complexity of the language used in candidates' websites or manifestos. Little attention has been heretofore given to discourse complexity in political economy (but see Spirling (2016)). This gap is surprising given that campaigns are, at their core, acts of communication, and politicians are often thought as experts at understanding their audience and finding the most efficient way to communicate with it. We may thus expect them to adjust the sophistication of their discourse depending on the voters they are trying to persuade. First, language complexity may send a signal of who the politician is and which group of voters they will represent once in office (e.g., "the people" vs. the elite). Second, voters may be more receptive to a campaign message that they can understand without being too simplistic, and different voters may prefer different levels of complexity. For instance, using a simple language may appeal to less educated voters as well as

[18] The final score of each document is further normalized as recommended by Martin and Vanberg (2007). See Appendix B.2 for more details.

[19] The magnitude of these correlations is in line with previous studies. For instance, Gentzkow, Shapiro and Taddy (2019) report a within-party correlation of 0.13 between their speech-based measure of ideology and representatives' DW-Nominate score.

immigrants whose native language is not English, while using a complex rhetoric may appeal to highly-educated voters.

One of the most widely used proxies for textual complexity is the Flesch–Kincaid readability metric. However, Benoit, Munger and Spirling (2019) found that more recent metrics brought by advances in software tools perform closer to human ratings when it comes to measuring different dimensions of textual sophistication in the field of politics. We combine their findings with the approach of Tolochko and Boomgaarden (2018) and Hurka and Haag (2020) to define textual complexity along three main dimensions: the complexity of sentences, the complexity of words, and the overall conceptual complexity.

First, we measure the syntactic complexity, which refers to the complexity of sentences' structure, by computing the ratio of the number of subordinating conjunctions and relative pronouns to the total number of words (Montoro and McIntyre, 2019; Benoit, Munger and Spirling, 2019).[20]

Second, we measure the semantic complexity, which refers to the complexity of words themselves. We approximate words' complexity by measuring their entropy, using Google Books as reference text (Michel and Orwant, 2011).[21] For any word appearing with frequency f in the Google Books corpus, we define its entropy as $-f \log f$. As a result, the entropy is mostly a decreasing function of words' frequency: the rarer a word, the higher its entropy. However, below a certain threshold, the entropy is an increasing function of words' frequency. This ensures that extremely rare words found in candidate websites and manifestos (such as typos or OCR errors) are not considered complex words.[22]

Third, we measure the conceptual complexity of texts, which refers to the difficulty to understand the ideas embedded within a text, regardless of its form. Levy, Razin and Young (2022) use the ratio of unique words to total words in speeches as a proxy for the complexity of ideas: more diverse words point to a more complex idea. This metric is also known as Type-Token Ratio (TTR), but it is strongly correlated with text length (e.g., a 10,000 word text will necessarily use the same words several times and will therefore have a lower ratio of unique words as compared to a very short and

[20] Some manifestos were processed through optical character recognition (OCR), which may limit our ability to identify punctuation. Instead of relying on sentence length, we count subordinating conjunctions such as "when," "where," "whether" in English and "que," "quoi," "où" in French. In practice, we use the R implementation of the OpenNLP package to identify subordinating conjunctions (part-of-the-speech tagging) in the U.S. Since this package does not support French language, we use lists from "Le Robert" dictionary to detect subordinating conjunctions and relative pronouns in French manifestos: https://dictionnaire.lerobert.com/guide/conjonctions-de-subordination and https://dictionnaire.lerobert.com/guide/pronoms-relatifs.

[21] We use the latest Google Books corpus available: 2008 in English and 2009 in French.

[22] Historically, semantic complexity was measured by the average number of syllables per word (Flesch, 1948) or the proportion of "difficult" words, i.e., words that are not included in a list of most common words (Dale and Chall, 1948). Benoit, Munger and Spirling (2019) use the average word rarity as given by the Google books corpus, an approach that is not immune to extremely rare words (likely to be typos or OCR errors). The measure of word entropy which we rely on, like Hahn and Sivley (2011) and Katz (2013), is analogous to Shannon's entropy in information theory (Shannon, 1948), which measures the computational resources required to process a telecommunication.

simpler text). To mitigate this problem, we use the Moving Average Type Token Ratio (MATTR), which computes the average TTR through a moving window of 200 words (Covington and McFall, 2010).[23]

We define the complexity score of a document as the standardized average of these three components: the share of subordinating words and relative pronouns, the average word entropy, and the MATTR.[24] We discuss the validity of this approach in Appendix B.3 and compare the complexity of our corpus against some benchmarks. We find that the average U.S. website is equally complex as the business/financial section of the New York Times, but less complex than its book review section and more complex than its sports section (Appendix Figure B.1). The 2022 French manifestos are more complex than their "easy to read and understand" counterparts, which candidates had the possibility to publish online along with their original manifesto in that specific year (Appendix Figure B.2).

3.3 Topic distribution

Besides their ideological tone and the complexity of their rhetoric, candidates may strategically choose which topics to focus on to persuade voters. To quantify the relative importance of different topics, we implement a supervised machine learning model trained on the manifestos issued by national parties.

For the U.S., we rely on the Manifesto Project (Lehmann et al., 2021), which gathered manifestos for the 2004-2020 period and hand coded each sentence to fit into one of 31 topics such as "Human rights," "Protectionism," "Education," or "Agriculture and farmers." For France, the party manifestos hand coded by the Manifesto Project are only available for the 2012 and 2017 elections. Therefore, we use an alternative data source: the French Agenda Project (Grossman, 2019), which gathered party manifestos for the 1981-2017 period and hand coded each sentence to fit into 27 topics such as "Economic regulation," "Health," "Education," or "Immigration." Appendix Table B.3 provides the full list of topics in each country, together with the words that are most predictive of each topic.

We feed the sentences from the national manifestos into a TF-IDF vectorizer and train a Support Vector Machine (SVM) model to predict each sentence's topic.[25] Once trained on the national manifestos, we used the SVM model to predict the topics appearing in our corpus. For each document,

[23] We use the Quanteda package (Benoit et al., 2018) to measure MATTR.

[24] For the complexity analysis, we cannot concatenate all the captures of a U.S. candidate's website together like we do for the ideological score, because some metrics such as the Type-Token Ratio are not linear. Instead, we define the complexity score as the average complexity of all website captures.

[25] This method yields a higher out-of-sample cross-validated accuracy than the alternative classifiers that we tried. We explain at greater length why we chose this approach in Appendix B.4.

the model outputs a vector indicating the likelihood that each topic is discussed.[26]

To check the quality of our topic predictions, we regress each measure of topic prevalence on our ideological score (Appendix Figure B.3) and complexity score (Appendix Figure B.4). On average, controlling for year fixed effects, candidates more to the left talk more about education and equality, while candidates more to the right focus on markets and administrative efficiency. More complex candidates speak more of international affairs, while less complex candidates refer more to labour groups.

3.4 Text similarity

Finally, we construct a measure of textual similarity between two documents based on their overall content. This measure is based on the average cosine similarity between the vector representations of texts, across multiple choices of such representations – including TF-IDF, Word2Vec, and BERT. Regardless of the technique chosen to transform texts into vectors, textual similarity relies on an unsupervised approach and is therefore more "agnostic" than measuring the distance between two texts' ideological scores, complexity scores, or topic distributions. It captures whichever parts of the text make two documents similar or different from each other, including other important dimensions that we could miss by just focusing on ideology, complexity, and topics. See Appendix B.5 for more details on this measure.

4 Convergence Between Rounds

We now investigate whether candidates adjust the content of their campaign communication to match their voters' preferences. We begin by computing correlations between voter characteristics and candidates' ideology and complexity scores in the second round.[27]

As shown in Appendix Tables C.1 and C.2, candidates tend to use more right-wing language in low-density districts (a correlation that is only significant in France) and in districts with stronger support for right-wing candidates in the previous presidential election (which is significant at the 1% level in both countries). Candidates running in more educated districts tend to use more left-wing language (significant at the 10% and 1% level in the U.S. and France, respectively) and more complex language (an estimate that is significant only in the U.S., but sizeable in France as well).

[26] Similarly as for the ideological score, in the U.S., we concatenate all the captures of a candidate's website prior to the day of the primary election on the one hand, and all the captures of their general election website between the primary and the general election on the other hand.

[27] See Appendix A.3 for more details on the voter characteristics we include in this analysis.

These correlations suggest that candidates adapt the content of their communication to their electorate, but they could be confounded by reverse causality (e.g., if more educated voters sort where more complex politicians run) or omitted variable bias (e.g., if other factors explain both why more educated voters and more complex politicians live in the same area).

To overcome these issues, we exploit the two-round setting of the U.S. and French elections and study how the same candidate adjusts their discourse to different electorates. Focusing on within-candidate changes in discourse between rounds enables us to hold many factors constant, including the political climate and candidates' identity. The fact that the same candidate targets different electorates in the first and second rounds is obvious in U.S. elections. As candidates who win the primary election move from competing within their own party to competing against the other party, their electorate broadens and becomes more diverse. In France, the set of eligible voters remains identical across rounds, but the composition of the actual electorate may change. For instance, extreme voters who participated in the first round may abstain from the second round if they feel too distant from any of the remaining candidates (Pons and Tricaud, 2018). Furthermore, the candidates who qualify for the runoff generally move from competing against many opponents in the first round, including other candidates from the same ideological side, to competing against a single candidate from the other side in the second round: 70% of runoff races in our sample oppose one left-wing and one right-wing candidate. Just like in the U.S., candidates have an incentive to target their base in the first round and to address a broader set of voters in the second round.

While candidates may be tempted to adjust their discourse to target a broader audience, flip-flopping may be costly, limiting the magnitude of the convergence. Indeed, since candidates anchored their platform in the first round, changing their discourse in the second round may hurt their reputation and cost them votes.

4.1 Ideological convergence

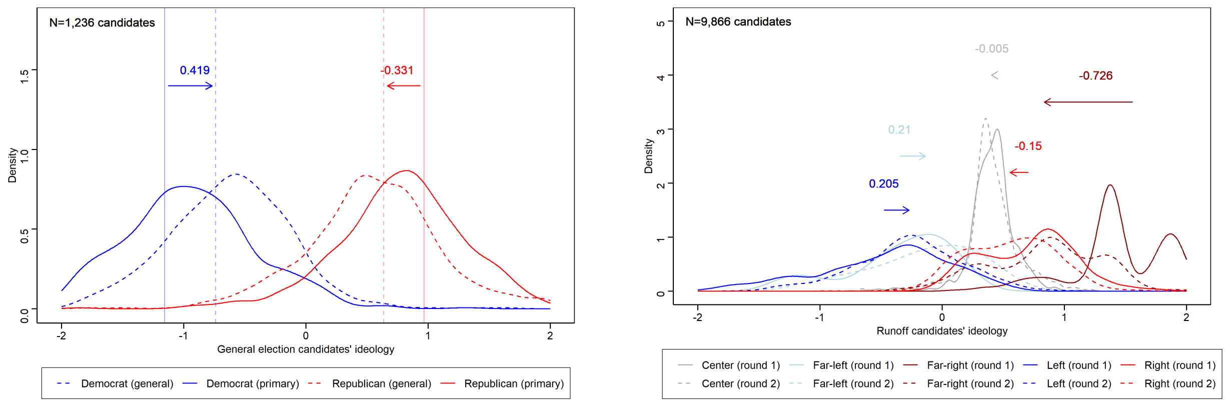
In Figure 1a, we plot the kernel density of ideological scores for Democratic and Republican candidates separately, pooling all elections year together. Our sample includes candidates who compete both in a competitive primary election (i.e., with two candidates or more) and in a competitive general election. The straight curves represent the distributions of ideological scores at the primary stage and the dashed curves represent the distributions of ideological scores among the same set of candidates at the general election stage.[28] Ideological scores are divided by their standard deviation

[28] As explained in Section 2.1, the website URLs of candidates present both in a competitive primary and a general election were collected through two different channels: the Wayback Machine search engine and the LoC, respectively. In this section, to ensure that we compare changes in ideological score from the same campaign websites (i.e., the same URLs) across election rounds, we focus on URLs taken from the LoC database and we use the captures of a candidates' general election website taken before the day of the primary election to calculate their primary ideological score. We use the captures of the same websites taken between the day of the primary and the general election to calculate the candidate's

at the primary stage.

We first observe that Democratic candidates tend to use left-wing language and Republican candidates right-wing language, both in the primary and the general election. This is somewhat mechanical since our method scores words primarily used by Democratic (Republican) candidates as left (right)-leaning. Second, both parties' distributions shift toward the center of the scale between the primary and general elections: the mean ideological score among Democrats shifts to the right by 0.42 standard deviation while the mean ideological score among Republicans shifts to the left by 0.33 standard deviation. Both estimates are significant at the 1% level, and they are not mechanical. These results indicate that the Democrats and Republicans that qualify for the general election tend to use more moderate language after the primary than before.

Figure 1: Ideology moderation



(a) U.S.

(b) France

Notes: We plot the kernel density of ideological scores for Democratic and Republican candidates in the U.S. (Figure 1a) and for the main political orientations in France (Figure 1b), pooling all election years together. The sample includes candidates who compete both in a competitive primary election and a competitive general election (Figure 1a), and candidates running both in a competitive first round and a competitive second round (Figure 1b). The solid curves represent the distributions of ideological scores in the first round and the dashed curves represent the distributions of ideological scores among the same set of candidates in the second round. In the U.S., candidates' ideological scores in the first round are calculated based on the captures of their general election website taken prior to the day of the primary election, while their scores in the second round are calculated based on the captures of their general election website between the primary and the general election. N=1,236 candidates (Figure 1a) and 9,866 candidates (Figure 1b).

Similarly, Figure 1b plots the kernel distribution of ideological scores among French candidates running both in a competitive first round and a competitive second round (pooling all election years together), for each political orientation separately. We observe that candidates from left-wing parties use more left-wing language than candidates from centrist parties, who generally fall on the right side of the scale, but not as far to the right as candidates from right-wing parties. In addition, candidates from the far-right tend to use more extreme language than candidates from the mainstream right. Interestingly, candidates from far-left parties are located closer to the center than left-wing candidates, on average, revealing that the very few far-left candidates who qualify for the

general ideological score.

runoff tend to be relatively moderate.[29] Unlike in the U.S., these patterns are not mechanical, since centrist candidates were excluded from the construction of word scores and candidates from the right and the far-right (respectively left and far-left) were pooled together. As in the U.S., the distributions of scores shift toward the center in the second round. This shift is visible for all orientations, and particularly strong for far-right candidates, who shift to the left by 0.73 standard deviation on average.[30]

In both countries, these patterns are robust to restricting the sample to second rounds with only two competing candidates (Appendix Figure C.1),[31] which is the setting closest to the assumptions of the median voter theorem.

4.2 Complexity convergence

In addition to moderating their ideological tone, candidates may adjust the complexity of their language to appeal to a broader set of voters. For instance, politicians who catered to a highly-educated voter base in the first round may simplify their discourse in the second round to appeal to less-educated voters.

In Figures 2a and 2b, we plot the kernel density of complexity scores for the same samples of candidates as for ideology. We show two separate distributions, for candidates whose complexity score in the first round is below vs. above the median.[32] In both countries, we find that the distributions shift toward the center of the complexity scale. We observe the same pattern when we restrict the analysis to second rounds with only two candidates (Appendix Figure C.2).

A possible concern is that these patterns could be partly driven by a reversion to the mean. However, Appendix Figure C.3 shows similar results when we predict candidates' complexity score in the first round (based on a regression of their actual complexity on their observable characteristics) and use this *predicted* complexity score to separate candidates in two groups.[33]

Finally, we test whether candidates' propensity to moderate their ideology and their complexity

[29] By contrast, among all first-round candidates, including those who do not qualify for a runoff, the average ideological score of far-left candidates is -1.34 standard deviations, against -0.63 for left-wing candidates.

[30] We note that Figure 1b replicates results from Le Pennec (2023) on a larger sample, since hers includes only parliamentary elections between 1958 and 1993 while ours also includes candidates running in parliamentary elections between 1997 and 2022 and in local elections between 1979 and 2021.

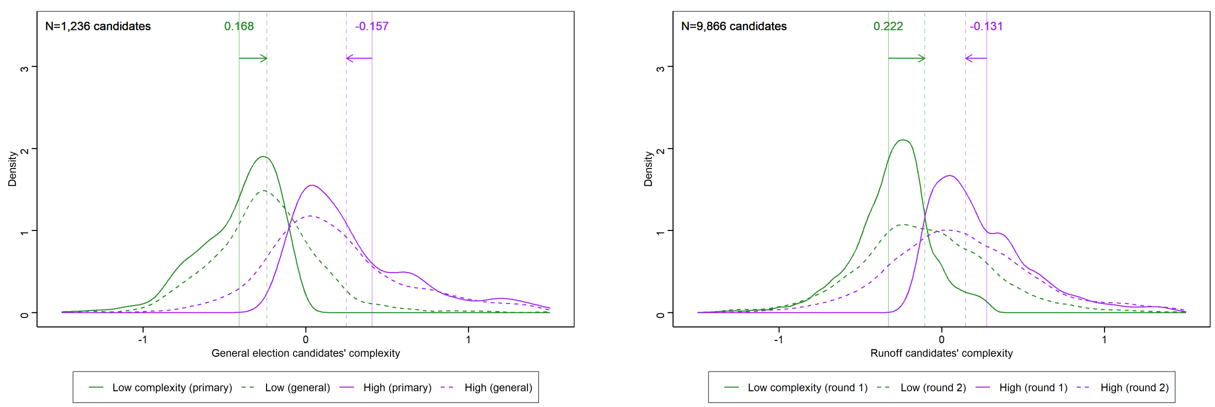
[31] More precisely, we exclude general elections where smaller independent candidates are present and where a primary election winner drops out before the general election (10 cases between 2002 and 2016), in the U.S.; and runoffs where more than two candidates are present in the second round as well as runoffs where two candidates qualify for the second round but one of them drops out of the race, in France.

[32] We compute the median complexity score in each election year separately.

[33] We predict complexity based on district fixed effects and candidate-specific variables: the candidate's party or political orientation, whether they are the incumbent, whether their party or political orientation won the previous election, their party or political orientation's vote share in the previous election, and the length of their website or manifesto.

are correlated. Appendix Figures C.4a and C.4b show a bin scatter plot of the mean complexity score against the mean ideology score in the first round as well as each bin's corresponding mean complexity and mean ideology in the second round. In the U.S., we observe that ideologically extreme candidates (either on the left or the right) tend to be more complex, whereas complexity does not vary as much in France. Despite these initial differences between countries, ideology and complexity tend to move toward the center in both settings, with larger adjustments for the most extremes candidates.

Figure 2: Complexity moderation



(a) U.S.

(b) France

Notes: We plot the kernel density of candidates' complexity score, pooling all election years together and splitting the sample between candidates whose complexity score in the first round is below the median score in a given election year, and those whose complexity score is above the median. Other notes as in Figure 1.

4.3 Topic convergence

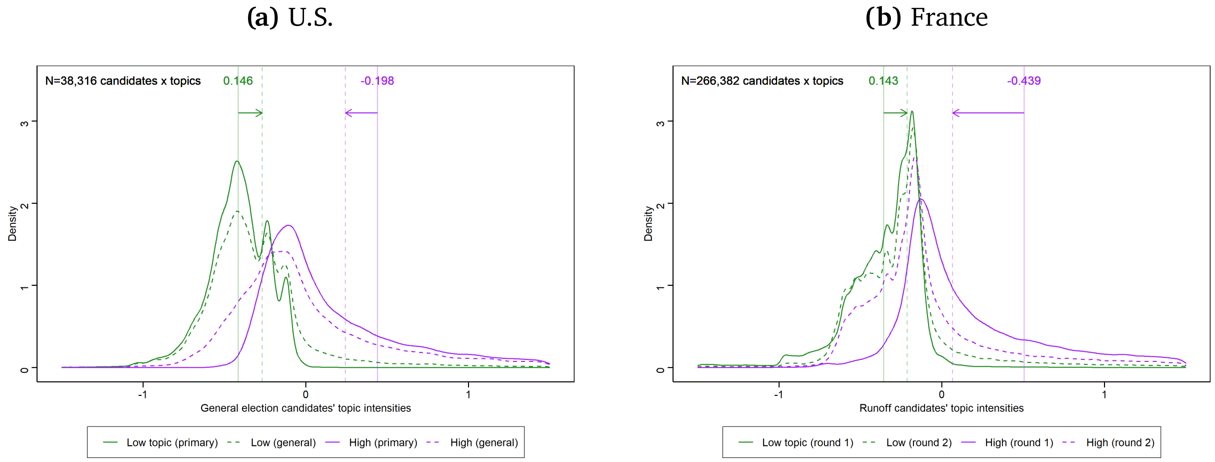
A third way for candidates to appeal to a broader electorate is to expand the set of topics they discuss. Candidates who campaigned on very specific topics to appeal to certain voters in the first round may give more space to other topics that other voters care about in the second round.

We assess the prevalence of each topic in the candidate's website or manifesto and plot the kernel density of topic prevalence in each election round, for high- and low-prevalence candidates separately, and pooling across all topics.[34] Figure 3 shows convergence to the center in both countries. The shift is particularly striking in France, where candidates who insisted a lot on some topics in the first round reduce the prevalence of these topics by 0.44 standard deviation in the second round. We observe the same pattern when we restrict the analysis to general elections and second rounds with only two candidates (Appendix Figure C.5) and when we separate candidates based on their

[34] Like for complexity, we distinguish high from low topic prevalence based on the median prevalence of that topic in the first round, in each election year separately.

predicted topic propensities in the first round (Appendix Figure C.6).[35]

Figure 3: Topics moderation



(a) U.S.

(b) France

Notes: We plot the kernel density of candidates' topic prevalence, pooling all election years and topics together. For each topic, we split the sample between candidates whose topic prevalence in the first round is below the median topic prevalence in a given election year, and those whose topic prevalence is above the median. N=38,316 candidates×topics (Figure 3a) and 266,382 candidates×topics (Figure 3b). Other notes as in Figure 1.

In sum, as their target electorate broadens and their number of competitors decreases, candidates move to the center of the ideology, complexity, and topics scales. In the next section, we show that candidates' convergence to the center is partly driven by their strategic convergence to their second-round opponent, which is the key mechanism underlying the median voter theorem.

5 Adjustment to Opponent

We now test whether candidates adjust their discourse to the opponent they are facing, exploiting races in which a candidate narrowly qualifies for the second round. We first provide two concrete examples of candidate adjustments.

In 2022, Franck Riester, a centrist candidate with an ideological score of 0.07, arrived first in the first round of parliamentary elections in the French *département* of Essonne. In the second round, he competed against François Lenormand, a far-right candidate with an ideological score of 0.93 who had narrowly qualified for the runoff after ranking second ahead of Cédric Colin, a left-wing candidate with an ideological score of -0.86. With Colin out of the race, Riester could reasonably expect left-wing voters to vote for him even if he did not make any effort to cater to them. Right-wing voters were more likely to hesitate between him and Lenormand. Moving his platform closer to Lenormand would help Riester persuade them and increase his vote share. Indeed, Riester shifted his manifesto to the right between the first and second round, reaching an ideological score of 0.30.

[35] We use the same procedure and the same variables to predict topic propensities as we did for complexity.

Specifically, he started using right-wing words such as "community" and "defense" in the runoff, and stopped using left-wing words such as "minimum pension," "climate emergency," or "income tax." This closed the gap between his initial ideological score and that of his far-right competitor while increasing the gap with the left-wing candidate eliminated after the first round.

We now turn to a U.S. example related to candidates' complexity. In 2016, in the fourth congressional district of Tennessee, Steve Reynolds won the Republican primary with a complexity score of 0.27. In the general election, he faced Scott Desjarlais, who had narrowly won the Democratic primary with a complexity score of -0.08 against Grant Starrett, whose complexity score was 0.55. After the primary, Reynolds updated his website and decreased his complexity to 0.06, thus closing the gap with the complexity of his opponent as opposed to the complexity of the Democratic runner-up. He used less complex words (-0.41 standard deviation), fewer subordinates (-0.19 standard deviation), and slightly less diverse words (-0.04 standard deviation).

5.1 Regression discontinuity design

Design We now test more systematically whether candidates adjust their discourse to that of their competitor by estimating the impact for a candidate – the *leader* – to face a certain contender – the *opponent* – in the second round of the election, instead of another potential contender – the *runner-up*.

In France, remember from Section 2.2 that the set of candidates who qualify for the second round includes the two candidates who received the most votes in the first round and any other candidate who obtained the votes of at least 12.5% of the registered voters. We call the candidate ranked first in the first round the *leader*. We focus on races in which the vote shares of the second and third candidates were very close to each other and lower than the 12.5% qualification threshold, such that the second candidate qualifies for the second round but the third is eliminated. We call the second candidate the *opponent* and the third candidate the *runner-up*.

In the U.S., we exploit close primary elections in which the two top candidates obtained nearly exactly the same vote shares. We call the winner of the primary the *opponent* and the second candidate the *runner-up*. Each primary (Democratic or Republican) is linked to a *leader*, defined as the candidate of the opposing party in the general election.[36] For instance, the leader associated with a Democratic primary is either the Republican nominee or a contender from a third party, when no Republican candidate runs in the general election. Some primary races cannot be linked to a leader and are excluded from the sample, including primaries whose winner will run unopposed in the general election and elections in which several third party contenders run in the general election

[36] Note that the leader is not necessarily the strongest candidate. We use this terminology by symmetry with the French context, in which the leader is the leading candidate in the first round.

but no Republican candidate does.

Our outcome, $Y_{i,l}$, is a measure of discourse convergence between the leader l and opponent or runner-up i in the second round. If the leader strategically adjusts their platform to their opponent in order to attract undecided voters, we should expect the convergence between them to be stronger than between the leader and the runner-up. In general, this pattern could also emerge absent strategic adjustment. Indeed, the leader may decide to emulate their opponent's platform because they reason that this platform appealed to more voters than the runner-up's platform in the first round. Our RDD rules out this confounding mechanism by focusing on races in which the first round vote shares of the opponent and the runner-up were nearly identical and their platforms can thus be expected to be equally appealing to voters.

Formally, our design uses two observations per race, measuring the convergence $Y_{i,l}$ between the leader and the opponent and the convergence between the leader and the runner-up, respectively. We define the running variable X as the difference between the vote shares obtained by the opponent and the runner-up in the first round. We set X as positive for the observation corresponding to the opponent, and negative for the observation corresponding to the runner-up. The treatment variable T is a dummy equal to one for the opponent ($X > 0$) and 0 for the runner-up ($X < 0$).[37]

We use a sharp regression discontinuity design and estimate the following equation:

$$Y_{i,l} = \alpha + \tau\, T_i + \beta\, X_i + \gamma\, T_i \times X_i + \epsilon_i. \tag{1}$$

We follow Imbens and Lemieux (2008) and Calonico, Cattaneo and Titiunik (2014) and estimate this specification non-parametrically, fitting a local linear regression on each side of the threshold within an optimal bandwidth selected by the MSERD procedure from Calonico et al. (2019). We cluster standard errors by district × year. We report robust p-values (Calonico, Cattaneo and Farrell, 2020).[38] Our coefficient of interest, τ, represents the causal effect, for the leader, of facing opponent i instead of the runner-up in the second round.

We define the outcome $Y_{i,l}$ by computing the distance between the leader's discourse in the first round and candidate i's discourse (also in the first round), the distance between the leader's discourse in the second round and candidate i's discourse (still in the first round), and taking negative the difference between them:

[37] In 17 races in France and one in the U.S., the opponent and the runner-up obtained the exact same numbers of votes. We exclude these races from the sample.

[38] We use the R implementation of the rdrobust package (Calonico et al., 2017).

$$Y_{i,l} = -\left(\underbrace{\left|Y_l^{(2)} - Y_i^{(1)}\right|}_{\substack{\text{Distance between leader } l \\ \text{at 2nd round and} \\ \text{opponent } i \text{ at 1st round}}} - \underbrace{\left|Y_l^{(1)} - Y_i^{(1)}\right|}_{\substack{\text{Distance between leader } l \\ \text{at 1st round and} \\ \text{opponent } i \text{ at 1st round}}} \right)$$

$Y_{i,l}$ takes a positive value if the leader moves their discourse toward candidate i's initial position between the first and second rounds.

Our design differs from standard close-election RDDs in two important ways. First, instead of using only one observation per constituency, with some constituencies falling above a threshold and others below, we use two observations per constituency, corresponding to the candidate above the qualification threshold (the opponent) and the candidate below (the runner-up). Second, close-election RDDs generally raise concerns of interpretation and external validity. By contrast, in our design, focusing on close elections is not just useful for identification. It enables us to compare a leader's adjustment to two potential opponents, in a setting in which these opponents and their discourse did equally well in the first round, so that the only difference between them is that one is present in the second round and the other is not.

Sample Our sample includes all races in which we observe the opponent or runner-up's discourse in the first round and the leader's discourse in both rounds. More precisely, in France, our sample includes races where both the first and second-round manifestos of the leader, and the first-round manifesto of either the opponent or the runner-up are available. In the U.S., we measure the first-round position of the leader based on captures of their general election website taken before the day of the primary election; and their second-round position based on captures of the same website between the primary and the general election. Remember from Section 2.1 that we used the LoC database to collect the URLs of general election websites. Since the runner-up does not qualify for the general election and is not included in this database, we measure the first-round position of both the runner-up and the opponent based on captures of their primary election websites, whose URLs were collected through the Wayback Machine's search engine. Therefore, our U.S. sample includes races in which the general election website of the leader and the primary election website of either the opponent or the runner-up are available.

In the U.S., if both the Democratic and Republican parties hold competitive primary races, a unique election and its two primary races can yield four observations, corresponding to the convergence between the Democratic (Republican) leader and their opponent and runner-up in the Republican (Democratic) primary. Since we cluster standard errors by district × year, these four observations are included in the same cluster. In both countries, there are races in which we observe the opponent's discourse but not the runner-up's, or vice versa, yielding one observation instead of two.

Overall, our sample includes a total of 1,852 observations across 1,225 races in the U.S., and 1,409 observations across 807 races in France. Appendix Tables D.1 and D.2 show the number of races along with the average number of candidates and qualifying margin in the first round. The qualifying margin is 26 percentage points on average in U.S. primary elections and 3.4 percentage points in French first-round elections.

In principle, our analysis could be affected by endogenous sample selection. A first concern is if the first-round manifesto or primary website of opponents qualified for the second-round is observed more often than that of runner-ups. Column 1 of Appendix Table D.3 shows that this is not the case: there is no significant jump in the probability of having a first-round manifesto or website available at the qualification threshold, in either country.

A second important concern is if the leader's decision to compete in the second round (instead of dropping out of the race) depends on the identity of the opponent. For instance, the leader may decide to stay in the race if the opponent they will face in the second round is extreme, and they may instead drop out if the opponent is moderate and thus deemed very likely to win. The latter type of races would be excluded from the sample, since we would not observe the leader's discourse in the second stage of the election. Fortunately, these cases are extremely rare and unlikely to affect our results. In France, no leader in our RDD sample ever drops out between the first and second rounds. In the U.S., only four primary election winners dropped out before the general election.

In Appendix E.1, we discuss these and other situations that could create endogenous sample selection at greater length, and provide empirical evidence that they are not a concern.

Outcomes Our main outcome $Y_{i,l}$ is defined as the *overall* change in similarity between leader l and opponent or runner-up i, which aggregates standardized changes in vectorized text similarity as well as changes in similarity in ideology, complexity, and topic distribution. We also consider changes in similarity along each dimension separately.

Vectorized text and topic distribution are multidimensional vectors. Therefore, changes in the leader's discourse between election rounds on these dimensions will impact their distance with the opponent and the runner-up differently.[39] To measure convergence in text similarity, we define the

[39] The leader's movements will only reduce the gap with the opponent and the runner-up in equal amounts if they all fall on the same line in the n dimension space and if the leader's position in the first and second round is either to the right or to the left (not between) the opponent and the runner-up. When $n = 1$, all three candidates are necessarily aligned. When $n > 1$, the leader will in general not be aligned with the opponent and the runner-up. Indeed, in a coordinate base including the opponent and the runner-up, this would mean that the leader's coordinate is a perfect 0 along the orthogonal dimension to the hyperplane formed by them.

outcome as the change in cosine similarity between the leader l and the opponent or runner-up i's text vectors, divided by its standard deviation and averaged across different vector representations of text (see Appendix B for more details). To measure convergence in topics, the outcome is negative the change in Euclidean distance between leader l and opponent or runner-up i's vectors of topic prevalence (as defined in Section 3.3).

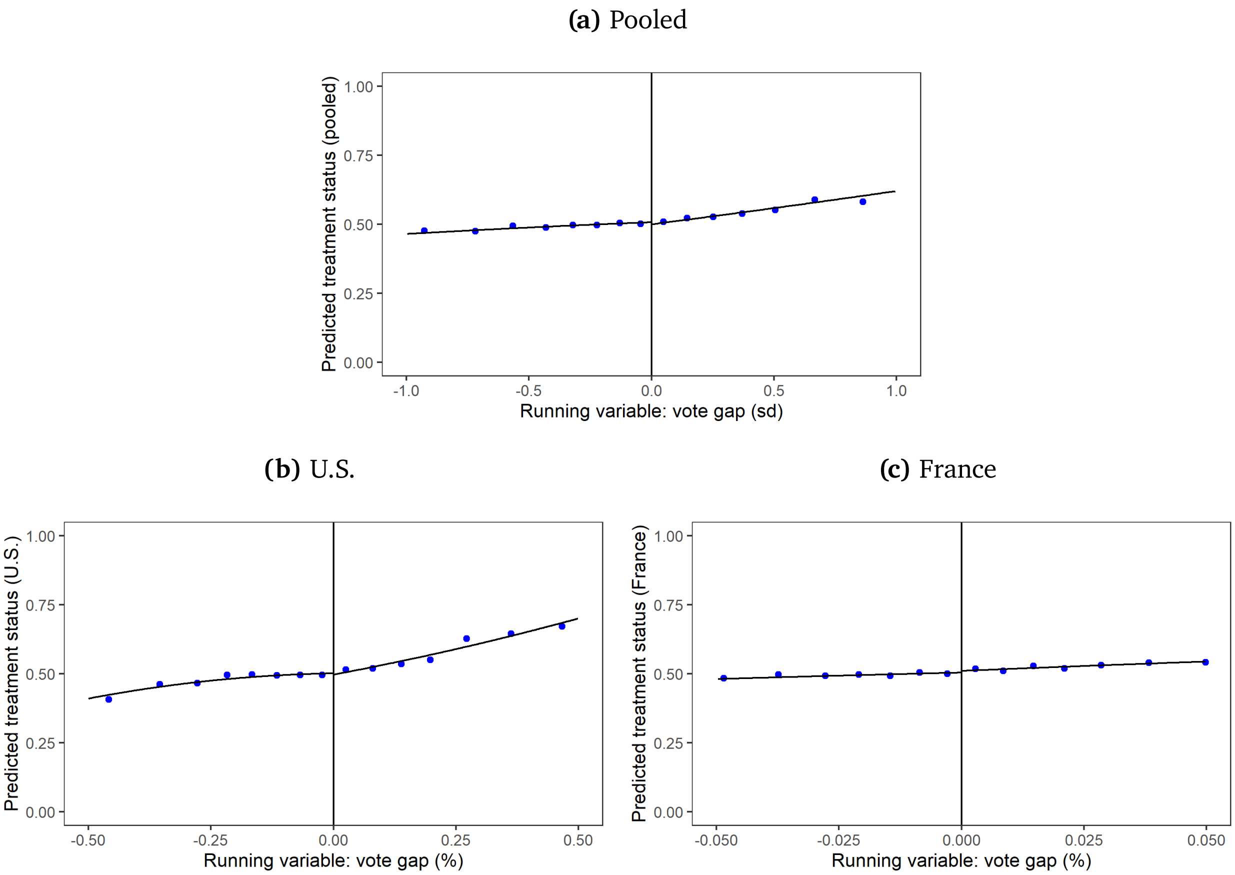
By contrast, the ideological score and the complexity score are both scalars. This can make comparing convergence to the opponent and to the runner-up challenging. If the leader is on the right of both the opponent and the runner-up in the first round and remains on their right in the second round, the change in similarity to both candidates will be exactly identical even if the leader tries to strategically adjust to the opponent. Therefore, when measuring convergence along these two outcomes, our main specification will restrict the sample to "middle point races": races in which the leader is initially between the opponent and the runner-up, either on the ideology or the complexity scale. In such races, changes in the leader's discourse will affect their distance to the opponent and the runner-up in opposite ways.[40] For ideology, we set the outcome as negative the change in distance between leader l and opponent or runner-up i's ideological scores (as defined in Section 3.1). For complexity, the outcome is negative the change in distance between the candidates' values of each complexity measure, divided by its standard deviation and averaged across the three measures of complexity (as defined in Section 3.2).

Identification assumption The validity of our RDD relies on the key assumption that first-round candidates of a certain type (e.g., candidates who are more similar to the leader) do not systematically sort on the right of the qualification threshold. Such manipulation is unlikely since it would require predicting the outcome of the first election stage with great accuracy, and unpredictable events make electoral outcomes uncertain.

We conduct several tests to bring empirical support for our identification assumption. First, we implement the test proposed by McCrary (2008) and verify that there is no discontinuity in the density of the running variable at the threshold. In our setting, this test would be satisfied by construction if data were available for all candidates, since we would have exactly two observations per race: one to the right of the threshold (the qualified opponent), and one to the left (the runner-up). Since our sample includes races for which we were able to collect the website or manifesto of one candidate but not the other, the McCrary test does not mechanically pass and remains informative. Appendix Figure D.1 shows that the null hypothesis of no sorting at the threshold cannot be rejected at conventional significance levels, whether in the U.S., in France, or when pooling the two samples together. Note that in this graph and in all analyses pooling both samples, we divide the running

[40] Appendix Table D.7 shows that there is no systematic candidate sorting across the qualification threshold in the subsample of middle point races.

Figure 4: General balance tests


(a) Pooled

(b) U.S.

(c) France

Notes: Dots represent the local averages of the predicted treatment status (vertical axis). Averages are calculated within quantile bins of the running variable (horizontal axis). The outcome is the candidate's predicted treatment status based on observable characteristics listed in the text. The treatment variable is a dummy equal to 1 if the candidate qualifies for the second round. The sample is restricted to candidates included in the RDD sample as described in the text. In Figure 4a, both countries' samples are pooled together and the running variable is divided by its standard deviation within each sample. In Figure 4b, the running variable is the vote share difference between the first two candidates in the primary election, and it is measured in percentage points. In Figure 4c, the running variable is the vote share difference between the second- and third-ranked candidates in the first round, and it is measured in percentage points. Continuous lines are a quadratic fit.

variable by its standard deviation in each country before pooling them together, in order to account for the fact that vote margins in France are tighter than in the U.S. Hence, the pooled running variable is measured in standard deviations.

Next, we conduct a general balance test by checking whether candidates' *predicted* treatment status jumps at the threshold. To predict treatment status, we first regress actual treatment on the following variables: a set of dummies indicating if the candidate is a Democrat (in the U.S.) or on the left (in France), if they are a woman, if they ran in the previous election in the same constituency, if they won that election, if their party or orientation won that election; the number of tokens in their website or manifesto; and their text similarity, ideological similarity, complexity similarity, and topic similarity to the first-round platform of the leader.

Table 1: General balance tests

Sample	Pooled (1)	U.S. (2)	France (3)
Treatment	0.005	0.017	0.015
	(0.013)	(0.023)	(0.016)
Robust p-value	0.604	0.324	0.344
Observations left	1509	821	688
Observations right	1752	1031	721
Effective obs. left	855	356	427
Effective obs. right	881	375	440
Polyn. order	1	1	1
Bandwidth	0.587	0.205	0.045
Mean, left of threshold	0.481	0.451	0.484

Notes: Standard errors, shown in parentheses, are clustered by district × year. We compute statistical significance based on the robust p-value and indicate significance at 1, 5, and 10% with ***, **, and *, respectively. The unit of observation is the candidate. The sample is restricted to candidates included in the RDD sample as described in the text. The outcome is the candidate's predicted treatment status based on observable characteristics listed in the text. In column 1, both countries' samples are pooled together and the running variable is divided by its standard deviation within each sample. We use local polynomial regressions: we fit separate polynomials of order 1 on each side of the threshold, using optimal bandwidths from the MSERD procedure. The mean, left of the threshold gives the mean value of the outcome for the runner-ups at the threshold.

Figure 4 shows the results. Each dot represents the probability of being treated within a given bin of the running variable – i.e., the vote share difference between the qualified opponent and the defeated runner-up. Opponents are located to the right of the threshold, and runner-ups to the left. A quadratic fit on each side of the cutoff is provided as a visual assistance. Figure 4a does not show any discontinuity in the full sample pooling U.S. and French observations. This is confirmed by the point estimate shown in column 1 of Table 1, which is small and non-significant. Figures 4b and 4c do not show any jump in the predicted probability of being treated at the qualification threshold either, in the U.S. or the French sample taken separately, and the corresponding point estimates are not significant (Table 1, columns 2 and 3).

We also test whether there is a discontinuity in any of the individual variables used to predict treatment. Appendix Tables D.3 through D.5 show the results, both for the U.S. and France. All estimates are small and non-significant, except for the probability that the same political orientation – either the candidate themselves or another candidate from that orientation – won the previous election in France, which is significant at the 1% level (Appendix Table D.4, Panel b). Given the large number of tests that we conduct (12 per country, for a total of 24), finding one coefficient significant at the 5% level would be in line with what we would expect. Finding one coefficient significant at the 1% level is more concerning. Reassuringly, our main results are similar when controlling for this covariate (Appendix Table D.6).

Finally, Appendix Figure D.2 shows that there is no discontinuity in the overall similarity between the candidate's and the leader's discourse in the first round. This provides reassuring evidence that the leader does not initially use language that is systematically more similar to the candidate who ends up qualifying as their opponent, as compared to the runner-up.

5.2 Main results

We first test for convergence to the opponent by measuring the impact, for a leader, of facing that opponent as opposed to their runner-up on the between-round change in overall similarity, which aggregates standardized changes in vectorized text similarity as well as changes in similarity in ideology, complexity, and topic distribution. Figure 5 shows that this outcome jumps up at the threshold in the U.S., in France, and in the pooled sample, indicating that leaders close the gap with their actual opponent more than with the runner-up who did not qualify. Table 2 complements the graphical analysis with formal estimates of the effects. Pooling both countries together, leaders' convergence to their actual opponent is 0.36 standard deviation larger than their convergence to the runner-up, which is significant at the 1% level. The effect is equal to 0.41 standard deviation and significant at the 1% level in the U.S. (column 2), and it is equal to 0.30 standard deviation and significant at the 5% level in France (column 3).

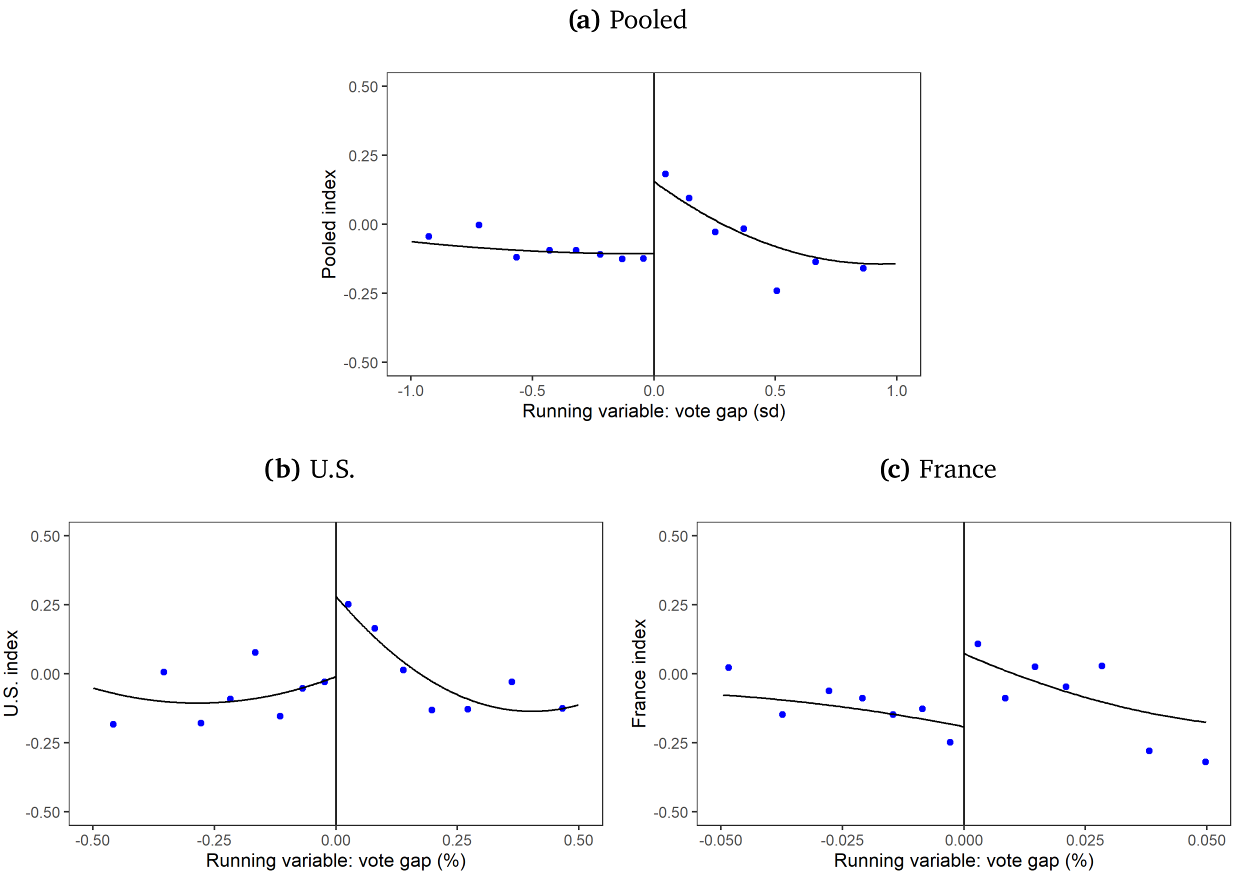
These results are robust to using a wide range of bandwidths (Appendix Figure D.3). They are similar if we restrict the sample to races in which exactly two candidates compete against each other in the second round, which is the setting closest to the assumptions of the median voter theorem (Appendix Table D.8).[41]

Next, we test for convergence to the opponent on each of the four textual dimensions separately to determine which of these dimensions drive the overall effect. The graphical results are shown in Appendix Figures D.4 and D.5 and the estimates are reported in Table 3.

Panel a shows the effects in the U.S. The effect on the change in vectorized text similarity is positive, sizeable (0.41 standard deviation), and significant at the 5% level (column 1), indicating that the leaders adjust their website to use language that is more similar to the language of their actual opponent, as compared to their opponent's runner-up. By contrast, the effect on ideological convergence is positive but small and non-significant (column 2). Importantly, note that in the U.S., we cannot estimate an effect for the subsample of middle point races by ideology, since there are only 19

[41] In France, we restrict the sample to races where exactly two candidates are present in the runoff, excluding races where either the leader or the qualified opponent drops out. In the U.S., we restrict the analysis to races where only the leader and the qualified opponent are present in the general election, without any other contender. These restrictions could be a source of endogenous sample selection: the identity of the qualified opponent may determine whether that candidate stays in the race to face the leader in the second round or decides to drop out instead, thus affecting the likelihood of a second round featuring exactly two candidates. In the U.S., the identity of the primary winner may also determine how many candidates choose to compete in the general election against that opponent. We discuss and provide evidence against these concerns in Appendix E.1.

Figure 5: Overall convergence


(a) Pooled

(b) U.S.

(c) France

Notes: The outcome is the change in overall similarity to the opponent or runner-up between election rounds, defined as the average of the standardized changes in vectorized text similarity as well as similarity in ideology, complexity, and topic distribution. It is constructed separately and divided by its standard deviation within each country. Other notes as in Figure 4.

elections in which the leader's primary ideological score falls between the ideological scores of the opponent and the runner-up. Indeed, it is rare for a Republican to use more left-wing language than either of the two Democratic candidates, and vice-versa. This limitation reduces our ability to test for strategic ideological convergence in the U.S. By contrast, races in which the leader's first-round complexity is initially in the middle of the opponent and the runner-up are much more common. Using that subsample, we find that leaders' complexity convergence to their actual opponent is 0.46 standard deviation larger than their convergence to the runner-up, which is significant at the 10% level (column 4). This pattern is consistent across all measures of complexity taken separately, although not as precisely estimated (Appendix Table D.9). Using the sample of all races, including those where the leader's initial complexity is not between the opponent and the runner-up yields a slightly smaller estimate (0.35 standard deviation) that remains significant at the 5% level (column 3). Finally, we do not find any significant convergence in topic distribution (column 5). In sum, U.S candidates' convergence to their opponent is primarily driven by an adjustment of their linguistic complexity.

Table 2: Overall convergence

Sample	Pooled (1)	U.S. (2)	France (3)
Treatment	0.357***	0.414***	0.300**
	(0.105)	(0.153)	(0.139)
Robust p-value	0.001	0.010	0.049
Observations left	1509	821	688
Observations right	1752	1031	721
Effective obs. left	855	425	417
Effective obs. right	881	439	432
Polyn. order	1	1	1
Bandwidth	0.588	0.253	0.044
Mean, left of threshold	-0.079	-0.049	-0.116

Notes: The outcome is the change in overall similarity to the opponent or runner-up between election rounds, defined as the average of the standardized changes in vectorized text similarity as well as similarity in ideology, complexity, and topic distribution. It is constructed separately and divided by its standard deviation within each country. Other notes as in Table 1.

Panel b shows the effects in France. The effect on the change in vectorized text similarity is positive but smaller than in the U.S. (0.11 standard deviation) and non-significant (column 1). By difference with the U.S., the number of middle point races is sufficient to test for strategic convergence not just in complexity but also in ideology. Indeed, it is common for the ideology of the candidate ranked first in the first round to be between the ideology of the second and third candidates. Think for instance of a race in which these three candidates are in the center, on the left, and on the right, respectively. We find that French candidates' overall convergence to their opponent is primarily driven by convergence on ideology: while the point estimate is not significant for the full sample (column 2), restricting the sample to middle point races, in which we expect the largest effects, yields a large effect of 0.46 standard deviation, significant at the 1% level (column 3). By contrast, we do not find any evidence of convergence in complexity, including when restricting the sample to middle point races (column 5). Finally, column 6 shows a positive effect on convergence in topic distribution, equal to 0.26 standard deviation, and significant at the 10% level.

These results indicate that candidates in both countries are strategic. Rather than running purely on conviction, they adjust their discourse to their opponent. In the U.S., where it is harder to study convergence in ideology, politicians try to appeal to undecided voters by adjusting the complexity of their discourse toward that of their opponent without changing the topics they discuss. This specific type of strategic adjustment – on the style rather than the content – may result from the country's bipartisan setting and high level of polarization. In that context, changing topics may be perceived as flip-flopping and cost politicians votes, whereas adjusting their level of complexity may appeal to new voters without antagonizing the base. In the French multipartisan setting, ideological positions are more malleable and candidates adjust their ideological tone rather than their complexity to their

Table 3: Convergence on different dimensions

(a) U.S.

Outcome	Text similarity	Ideology	Complexity		Topics
	Full sample	Full sample	Full sample	Middle points	Full sample
	(1)	(2)	(3)	(4)	(5)
Treatment	0.412**	0.153	0.353**	0.455*	0.076
	(0.153)	(0.143)	(0.158)	(0.250)	(0.133)
Robust p-value	0.015	0.263	0.040	0.090	0.617
Observations left	821	821	821	207	821
Observations right	1031	1031	1031	207	1031
Effective obs. left	409	408	445	87	360
Effective obs. right	425	422	458	87	378
Polyn. order	1	1	1	1	1
Bandwidth	0.240	0.239	0.268	0.236	0.208
Mean, left of threshold	0.121	0.330	0.090	-0.064	-0.061

(b) France

Outcome	Text similarity	Ideology		Complexity		Topics
	Full sample	Full sample	Middle points	Full sample	Middle points	Full sample
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.112	0.205	0.463***	0.133	-0.042	0.261*
	(0.129)	(0.124)	(0.162)	(0.112)	(0.174)	(0.152)
Robust p-value	0.444	0.132	0.007	0.335	0.871	0.095
Observations left	688	688	312	688	172	688
Observations right	721	721	312	721	172	721
Effective obs. left	485	356	186	399	100	415
Effective obs. right	496	364	186	412	100	430
Polyn. order	1	1	1	1	1	1
Bandwidth	0.054	0.033	0.043	0.041	0.036	0.043
Mean, left of threshold	-0.250	0.230	-0.047	0.001	-0.104	-0.020

Notes: The outcome is the change in similarity to the opponent or runner-up between election rounds, in terms of text similarity (column 1, Panels a and b), ideological score (column 2, Panel a, and columns 2 and 3, Panel b), complexity score (columns 3 and 4, Panel a, and columns 4 and 5, Panel b), and topic distribution (column 5, Panel a, and column 6, Panel b). Each outcome is divided by its standard deviation. In column 3, Panel b, the sample is restricted to races in which the leader is initially in the middle of the opponent and the runner-up on the ideology scale. In column 4, Panel a, and column 5, Panel b, the sample is restricted to races in which the leader is initially in the middle of the opponent and the runner-up on the complexity scale. Other notes as in Table 2.

opponent. They also change the relative importance that they give to different topics to match the distribution of topics in their opponent's discourse more closely.

Importantly, note that the convergence remains incomplete. In fact, if candidates fully converged to the median voter, the leader would not have any incentive to converge to the first-round platform of the opponent. No matter who qualifies for the second round, the leader should always converge to the median, expecting that the opponent would do the same. Instead, our design enables us to demonstrate the existence of a *partial* convergence equilibrium. Strategic convergence may be constrained by, for instance, party discipline, the cost of being seen as flip-flopping, and credibility concerns. Since the leader does not expect the opponent to fully converge, getting closer to them may swing voters who would hesitate between the two candidates otherwise.

5.3 Heterogeneity analysis

We now investigate the sources of variation in the level of adjustment across candidates. First, if candidates respond to incentives, they may adjust more when their chances of victory really depend on it. Second, some types of candidates may be willing to adjust more than others, e.g., if they care about winning more than about defending their ideas.

To test the first hypothesis, we ask whether leaders converge more to their opponent when the latter poses a greater threat to them. We use an alternative RDD to estimate the effect of facing a more extreme opponent as opposed to a more moderate one. We identify the most extreme of the two candidates that the leader could be opposed to in the second round – e.g., the most extreme of the top two candidates in the Democratic primary that the Republican nominee could have to face – and define the running variable as the difference in vote shares between them and the other potential opponent. The running variable is positive (and the treatment equal to 1) in races where the qualified opponent is more extreme, and it is negative (and the treatment equal to 0) when the qualified opponent is more moderate. Unlike in our main RDD, there is a unique observation per race. We use the leader's between-round change in overall similarity to the qualified opponent as the outcome. By virtue of the RDD, leaders immediately to the left of the threshold can be expected to be comparable to those immediately to the right, except for the type of opponent that they face. If leaders believe that more extreme opponents have a smaller base and that they can attract fewer new voters in the second round, they may adjust less to them, which would yield a negative effect.

Table 4 shows the impact of facing a more extreme opponent, where extremeness is defined successively in terms of ideology and complexity. See Appendix Figure D.6 for the graphical evidence.[42] Although the effects are not statistically significant, they are negative and large in both countries, indicating that facing a more extreme opponent causes leaders to converge less to them. Specifically, convergence to the more ideologically extreme opponent is 0.46 standard deviation lower in the U.S. and 0.24 standard deviation lower in France (columns 1 and 3). We obtain similar results, although smaller in size, when we define opponent's extremeness based on their level of complexity (Table 4, columns 2 and 4).

Leaders' incentives to adjust may be higher not just when they face a moderate opponent but also when they face an incumbent or an opponent whose political orientation received more votes in the previous election. Using separate RDDs, we find that both of these treatments lead candidates to

[42] We also verify that there is no discontinuity in the density of the running variable (Appendix Table D.10) and in the predicted treatment status (Appendix Table D.11) at the threshold.

Table 4: Convergence to the more extreme opponent

Sample	U.S.		France	
	Extreme ideology	Extreme complexity	Extreme ideology	Extreme complexity
	(1)	(2)	(3)	(4)
Treatment	-0.455	-0.165	-0.236	-0.208
	(0.271)	(0.288)	(0.221)	(0.219)
Robust p-value	0.143	0.556	0.506	0.351
Observations left	327	320	372	309
Observations right	300	307	230	293
Effective obs. left	168	161	184	177
Effective obs. right	169	168	139	169
Polyn. order	1	1	1	1
Bandwidth	0.290	0.279	0.035	0.040
Mean, left of threshold	0.149	0.005	0.006	-0.057

Notes: The outcome is the change in overall similarity between the leader and the opponent between election rounds, defined as the average of the standardized changes in vectorized text similarity as well as similarity in ideology, complexity, and topic distribution. It is constructed separately and divided by its standard deviation within each country. There is one observation per race and the running variable is defined as the difference in vote shares between the more extreme candidate (either in terms of ideology or complexity) and the other potential opponent. It is positive in races where the qualified opponent is more extreme and negative in races where the qualified opponent is more moderate. Other notes as in Table 1.

converge more to their opponent, in French elections (Appendix Table D.12).[43] Although sample sizes are much smaller, the effect of facing the incumbent is significant at the 10% level (column 1). Overall, these results confirm that candidates respond to incentives: they are more likely to adjust their discourse when their opponent is a stronger candidate who is likely to gather many votes.

We now turn to our second hypothesis and investigate differences in our main RDD effects across different types of candidates. In Appendix Tables D.15 and D.16, we replicate our main results from Section 5.2 for different subsamples of leaders, using the overall convergence index as the outcome. Although the effects are not all significant, we observe strategic convergence to the opponent across all types of candidates. In particular, the effects are comparable in size among left-wing and right-wing candidates (below and above the median ideological score).

Nevertheless, we do notice interesting differences. In the U.S., convergence tends to be larger among leaders who already ran in the past and incumbents (Appendix Table D.15, Panel a, columns 3 through 6) as well as leaders running in a district where their party received a large support in the previous election (Appendix Table D.16, Panel a, columns 3 and 4).[44] In France, it is larger among leaders who received a large number of votes in the first round (Appendix Table D.16, Panel

[43] Appendix Tables D.13 and D.14 show that there is no discontinuity in the density of the running variable at the threshold and in the predicted treatment status, for either treatment. We cannot run this analysis in the U.S. since too few incumbents run in a competitive primary and since both potential opponents are from the same party, making the vote share of their orientation in the previous election mechanically identical. In France, the sample size is limited because we need to restrict the sample to races where either the second- or third-ranked candidate is the incumbent (column 1) and races where the second- and third-ranked candidates are from different orientations (column 2).

[44] Large support is defined as receiving more than the median vote share in the previous general election.

b, columns 1 and 2). These patterns suggest that some candidates are consistently willing to adjust their discourse more than others, which contributes to their success ex ante, and explains why they continue to adjust more ex post.

In sum, the extent to which candidates adjust their discourse to their opponent reflects both their type and the electoral incentives that they face. We finally check whether candidates' propensity to strategically converge to their opponent has changed over time and find that it is larger today. Splitting the sample period in two halves reveals that the effect is already substantial before 2008 in both countries (Appendix Table D.17, columns 1 and 3), but it is larger and only statistically significant after 2008 (columns 2 and 4).

6 Conclusion

This paper provides the first direct empirical test of the convergence mechanism underlying the median voter theorem. Using a novel database containing 9,000 candidate websites for the primary and general election of U.S. House of Representatives between 2002 and 2016, as well as 57,000 candidate manifestos issued for the first and second round of French parliamentary and local elections between 1958 and 2022, we derive two sets of results.

First, we find that, as their target electorate broadens from the first to second round, candidates present in both stages adjust their discourse and "move to the center." They do so by moderating their ideological tone as well as the complexity of their discourse, and by diversifying the topics they talk about.

Second, we show that this convergence to the center often results from candidates adjusting to the rival they will face in the second round, once they learn their identity: politicians "follow each other." Our RDD exploits races in which the identity of the rival is quasi-random. In some circumstances, the rival emerges after a close contest between very different politicians, so that strategic adjustments to the winner have the opposite sign to adjustments that would be made in response to the platform presented by the loser. An example from France involves a centrist politician who ranked first in the first round and, in the runoff, must face a left-wing politician who barely qualified ahead of a right-wing candidate. On average, we observe that the centrist candidate moves to their left between the first and second rounds in that case.

In the U.S., the system of primaries prevents identifying ideological adjustment to the opponent as the two possible rivals are always on the same side of the ideological spectrum. For instance, the Republican nominee always has both possible rivals in the general election emerge on their left, from the Democratic primary, so closing the gap with the candidate who won the primary would also close it with the candidate who lost. Accordingly we study an alternative dimension to ideology: the complexity of the language used. In many races, a politician a) faces a rival that narrowly defeated their opponent in the primary, and b) has a measure of language complexity that is between that of the two possible contenders. Then, the politician will generally change the complexity of their discourse to match that of their opponent, simplifying it if the less complex opponent qualified, and moving in the opposite direction otherwise.

Pooling the U.S. and French samples and using an index that summarizes convergence in ideology, complexity, as well as topics and vectorized text, we find that the change in similarity to the qualified opponent is 0.36 standard deviation larger than the change in similarity to the runner-up.

Although politicians do not fully converge to the median voter, they do strategically get closer to each other. They may have their own ideas and convictions, but they certainly act on incentives to win elections nonetheless, just like Downsian models of electoral competition predict. Furthermore, strategic convergence is not limited to ideological tone and policy platform: politicians also adapt the complexity of their language. Beyond providing evidence on the strategies that candidates use to defeat their competitors, these results show that deciphering politicians' behavior requires studying not just the substance of their discourse but also its form.

7 References

Abou-Chadi, Tarik. 2016. "Niche party success and mainstream party policy shifts–how green and radical right parties differ in their impact." *British Journal of Political Science* 46(2):417–436.

Abou-Chadi, Tarik and Werner Krause. 2020. "The causal effect of radical right success on mainstream parties' policy positions: A regression discontinuity approach." *British Journal of Political Science* 50(3):829–847.

Acree, Brice D.L., Justin H. Gross, Noah A. Smith, Yanchuan Sim and Amber E. Boydstun. 2020. "Etch-a-Sketching: Evaluating the post-primary rhetorical moderation hypothesis." *American Politics Research* 48(1):99–131.

Adams, James. 2012. "Causes and electoral consequences of party policy shifts in multiparty elections: Theoretical results and empirical evidence." *Annual Review of Political Science* 15:401–419.

Adams, James and Samuel Merrill. 2003. "Voter turnout and candidate strategies in American elections." *The Journal of Politics* 65(1):161–189.

Adena, Maja and Anselm Hager. 2022. "Does online fundraising increase charitable giving? A nationwide field experiment on Facebook." *WZB Discussion Paper* (2020–302r).

Anagol, Santosh and Thomas Fujiwara. 2016. "The runner-up effect." *Journal of Political Economy* 124(4):927–991.

Ansolabehere, Stephen, James Snyder and Charles Stewart. 2001. "Candidate positioning in US House elections." *American Journal of Political Science* pp. 136–159.

Ash, Elliott, Massimo Morelli and Matia Vannoni. 2023. "More Laws, More Growth? Evidence from U.S. States." *CEPR Discussion Paper* (DP15629).

Benoit, Kenneth, Kevin Munger and Arthur Spirling. 2019. "Measuring and Explaining Political Sophistication through Textual Complexity." *American Journal of Political Science* 63(2):491–508.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller and Akitaka Matsuo. 2018. "quanteda: An R package for the quantitative analysis of textual data." *Journal of Open Source Software* 3(30):774.
URL: *https://quanteda.io*

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

Boche, Adam, Jeffrey B. Lewis, Aaron Rudkin and Luke Sonnet. 2018. "The new Voteview.com: preserving and continuing Keith Poole's infrastructure for scholars, students and observers of Congress." *Public Choice* 176(1/2):17–32.

Bouton, Laurent, Julia Cagé, Edgard Dewitte and Vincent Pons. 2022. "Small Campaign Donors." *NBER Working Paper* (30050).

Burden, Barry C. 2001. "The polarizing effects of congressional primaries." *Congressional primaries and the politics of representation* pp. 95–115.

Cagé, Julia, Caroline Le Pennec and Elisa Mougin. 2023. "Firm Donations and Political Rhetoric: Evidence from a National Ban." *American Economic Journal: Economic Policy* (Forthcoming).

Calonico, Sebastian, Matias D. Cattaneo and Max H. Farrell. 2020. "Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs." *The Econometrics Journal* 23(2):192–210.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell and Rocio Titiunik. 2017. "rdrobust: Software for regression-discontinuity designs." *The Stata Journal* 17(2):372–404.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell and Rocio Titiunik. 2019. "Regression discontinuity designs using covariates." *Review of Economics and Statistics* 101(3):442–451.

Calonico, Sebastian, Matias D. Cattaneo and Rocio Titiunik. 2014. "Robust nonparametric confidence intervals for regression-discontinuity designs." *Econometrica* 82(6):2295–2326.

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

Canes-Wrone, Brandice, David W. Brady and John F. Cogan. 2002. "Out of step, out of office: Electoral accountability and House members' voting." *American Political Science Review* 96(1):127–140.

Cascio, Elizabeth U. and Ebonya Washington. 2014. "Valuing the vote: The redistribution of voting rights and state funds following the voting rights act of 1965." *The Quarterly Journal of Economics* 129(1):379–433.

Cattaneo, Matias D, Michael Jansson and Xinwei Ma. 2018. "Manipulation testing based on density discontinuity." *The Stata Journal* 18(1):234–261.

Covington, Michael A. and Joe D. McFall. 2010. "Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR)." *Journal of quantitative linguistics* 17(2):94–100.

Dale, Edgar and Jeanne S. Chall. 1948. *A formula for predicting readablility*. Columbus, O.: Bureau of Educational Research, Ohio State University.

Dano, Kevin, Francesco Ferlenga, Vincenzo Galasso, Caroline Le Pennec and Vincent Pons. 2023. "Coordination and Incumbency Advantage in Multi-Party Systems - Evidence from French Elections." *NBER Working Paper* (30541).

Davis, Otto A. and Melvin J Hinich. 1968. "On the power and importance of the mean preference in a mathematical model of democratic choice." *Public Choice* pp. 59–72.

Davis, Otto A., Melvin J. Hinich and Peter C. Ordeshook. 1970. "An Expository Development of a Mathematical Model of the Electoral Process." *The American Political Science Review* 64(2):426–448.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *CoRR* abs/1810.04805.

Downs, Anthony. 1957. "An economic theory of political action in a democracy." *Journal of Political Economy* 65(2):135–150.

Druckman, James N., Martin J. Kifer and Michael Parkin. 2010. "Timeless Strategy Meets New Medium: Going Negative on Congressional Campaign Web Sites, 2002–2006." *Political Communication* 27(1):88–103.

Druckman, James N., Martin J. Kifer and Michael Parkin. 2018. "Resisting the Opportunity for Change: How Congressional Campaign Insiders Viewed and Used the Web in 2016." *Social Science Computer Review* 36(4):392–405.

Erikson, Robert S. and Gerald C Wright. 1980. "Policy representation of constituency interests." *Political Behavior* 2(1):91–106.

Fauconnier, Jean-Philippe. 2015. "French Word Embeddings.".
URL: *http://fauconnier.github.io*

Flesch, Rudolph. 1948. "A new readability yardstick." *Journal of Applied Psychology* 32(3):221–233.

Foarta, Dana and Massimo Morelli. 2021. "Complexity and the reform process." *CEPR Discussion Paper* (DP16562).

Fowler, Anthony. 2013. "Electoral and policy consequences of voter turnout: Evidence from compulsory voting in Australia." *Quarterly Journal of Political Science* 8(2):159–182.

Fujiwara, Thomas. 2015. "Voting technology, political responsiveness, and infant health: Evidence from Brazil." *Econometrica* 83(2):423–464.

Gaultier-Voituriez, Odile. 2016. "Archelec, les archives électorales de la Ve République, du papier au numérique." *Histoire@ Politique* (3):213–220.

Gentzkow, Matthew, Jesse M. Shapiro and Matt Taddy. 2019. "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech." *Econometrica* 87(4):1307–1340.

Gerber, Alan S. and Donald P. Green. 2000. "The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment." *American Political Science Review* 94(3):653–663.

Granzier, Riako, Vincent Pons and Clémence Tricaud. 2023. "Coordination and Bandwagon Effects: How Past Rankings Shape the Behavior of Voters and Candidates." *American Economic Journal: Applied Economics* (Forthcoming).

Grofman, Bernard. 2004. "Downs and two-party convergence." *Annual Review of Political Science* 7(1):25–46.

Grossman, Emiliano. 2019. "The French Agendas Project." *Comparative Policy Agendas: Theory, Tools, Data* p. 90.

Hahn, Lance W. and Robert M. Sivley. 2011. "Entropy, semantic relatedness and proximity." *Behavior Research Methods* 43(3):746–760.

Hall, Andrew B. 2015. "What happens when extremists win primaries?" *American Political Science Review* 109(1):18–42.

Hotelling, Harold. 1929. "Stability in Competition." *The Economic Journal* 39(153):41–57.

Hurka, Steffen and Maximilian Haag. 2020. "Policy complexity and legislative duration in the European Union." *European Union Politics* 21(1):87–108.

Husted, Thomas A. and Lawrence W. Kenny. 1997. "The Effect of the Expansion of the Voting Franchise on the Size of Government." *Journal of Political Economy* 105(1):54–82.

Imbens, G. and K. Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *The Review of economic studies* 79(3):933–959.

Imbens, Guido W. and Thomas Lemieux. 2008. "Regression discontinuity designs: A guide to practice." *Journal of Econometrics* 142(2):615–635.

Katz, Daniel Martin. 2013. "Measuring the Complexity of the Law: The United States Code." p. 41.

Koh, Allison, Daniel Kai Sheng Boey and Hannah Béchara. 2021. Predicting Policy Domains from Party Manifestos with BERT and Convolutional Neural Networks. preprint SocArXiv.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting policy positions from political texts using words as data." *American Political Science Review* 97(2):311–331.

Le Pennec, Caroline. 2023. "Strategic Campaign Communication: Evidence from 30,000 Candidate Manifestos." *SoDa Laboratories Working Paper Series* (2020-05).

Lee, David S., Enrico Moretti and Matthew J. Butler. 2004. "Do voters affect or elect policies? Evidence from the US House." *The Quarterly Journal of Economics* 119(3):807–859.

Lehmann, Pola, Theres Matthieß, Nicolas Merz, Sven Regel and Annika Werner. 2021. "Manifesto Corpus.". Version: 2021a.

Levitt, Steven D. 1996. "How do senators vote? Disentangling the role of voter preferences, party affiliation, and senator ideology." *The American Economic Review* pp. 425–441.

Levy, Gilat, Ronny Razin and Alwyn Young. 2022. "Misspecified Politics and the Recurrence of Populism." *American Economic Review* 112(3):928–962.

Lindbeck, Assar and Jörgen W. Weibull. 1987. "Balanced-budget redistribution as the outcome of political competition." *Public Choice* 52:273–297.

Martin, Lanny W. and Georg Vanberg. 2007. "A robust transformation procedure for interpreting political text." *Political Analysis* 16(1):93–100.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

McCrary, Justin. 2008. "Manipulation of the running variable in the regression discontinuity design: A density test." *Journal of Econometrics* 142(2):698–714.

Meltzer, Allan H. and Scott F. Richard. 1983. "Tests of a rational theory of the size of government." *Public Choice* 41(3):403–418.

Merz, Nicolas, Sven Regel and Jirka Lewandowski. 2016. "The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis." *Research & Politics* 3(2):2053168016643346.

Michel, Jean-Baptiste, Yuan Kui Shen Pinker Steven Nowak Martin A. Aiden Erez Lieberman Aiden Aviva Presser Veres Adrian Gray Matthew K. Pickett Joseph P. Hoiberg Dale Clancy Dan Norvig Peter and Jon Orwant. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science (American Association for the Advancement of Science)* 331(6014):176–182.

Miller, Grant. 2008. "Women's suffrage, political responsiveness, and child survival in American history." *The Quarterly Journal of Economics* 123(3):1287–1327.

Miller, Michael and Nicki Camberg. 2020. "U.S. House Primary Election Results (2012-2018).".

MIT Election Data and Science Lab. 2017*a*. "U.S. House 1976–2020.".

MIT Election Data and Science Lab. 2017*b*. "U.S. President 1976–2020.".

Montoro, Rocío and Dan McIntyre. 2019. "Subordination as a potential marker of complexity in serious and popular fiction: a corpus stylistic approach to the testing of literary critical claims." *Corpora* 14(3):275–299.

Nokken, Timothy P. and Keith T. Poole. 2004. "Congressional party defection in American history." *Legislative Studies Quarterly* 29(4):545–568.

OpinionWay. 2017. Les Français et les programmes électoraux. Sondage OpinionWay pour Le Printemps de l'Economie.

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

Palfrey, Thomas R. 1984. "Spatial equilibrium with entry." *The Review of Economic Studies* 51(1):139–156.

Persson, Torsten and Guido Tabellini. 2002. *Political economics: explaining economic policy*. MIT press.

Pettigrew, Stephen, Karen Owen and Emily Wanless. 2014. "U.S. House Primary Election Results (1956-2010).".

Pons, Vincent. 2018. "Will a five-minute discussion change your mind? A countrywide experiment on voter choice in France." *American Economic Review* 108(6):1322–63.

Pons, Vincent and Clémence Tricaud. 2018. "Expressive voting and its cost: Evidence from runoffs with two or three candidates." *Econometrica* 86(5):1621–1649.

Poole, Keith T and Howard Rosenthal. 1985. "A spatial model for legislative roll call analysis." *American journal of political science* pp. 357–384.

Shannon, Claude Elwood. 1948. *The mathematical theory of communication*. Urbana: University of Illinois Press.

Spenkuch, Jörg L. and David Toniatti. 2018. "Political advertising and election results." *The Quarterly Journal of Economics* 133(4):1981–2036.

Spirling, Arthur. 2016. "Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915." *The Journal of Politics* 78(1):120–136.

Tolochko, Petro and Hajo G. Boomgaarden. 2018. "Analysis of Linguistic Complexity in Professional and Citizen Media." *Journalism studies (London, England)* 19(12):1786–1803.

Underhill, Wendy. 2017. "Primary runoff elections.". Accessed: 2022-07-29.
URL: *https://www.ncsl.org/research/elections-and-campaigns/primary-runoff-elections.aspx*

Wittman, Donald. 1977. “Candidates with policy preferences: A dynamic model.” *Journal of economic Theory* 14(1):180–189.

Wittman, Donald A. 1973. “Parties as utility maximizers.” *American Political Science Review* 67(2):490–498.

Woolley, John T. and Gerhard Peters. 2017. “The American presidency project.” *Santa Barbara, CA. Available from World Wide Web: http://www. presidency. ucsb. edu/ws* .

Yamada, Ikuya, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji and Yuji Matsumoto. 2020. Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics pp. 23–30.

Online Appendix

Table of Content

A Additional Details on the Setting and Data

A.1 U.S. candidate websites

Finding primary election websites As indicated in Section 2.1, the websites' URLs of general election candidates were retrieved from the Library of Congress. The Library of Congress does not contain the websites' URLs of candidates defeated in the primary elections. Therefore, we collected all primary elections campaign websites' URLs (whether the candidate won or lost) using the search engine of the Wayback Machine.

The Wayback Machine is a non-profit initiative that aims to keep track of the Internet's most important content. It stores websites as they appeared at different points in time (see for example Figure A.1), which enables us to track how candidates' websites evolve during the campaign, from the primary to the general election. The Wayback Machine stores websites' main web page (the landing page) as well as web pages accessible through links on the main page. Hence, for each website, we have access to several web pages, and for each web page, several time captures.

The Wayback Machine's search function matches keywords entered by users with the URLs and titles of all archived web pages (more than 820 billion). To make sure we collected as many primary election websites as possible (in competitive primaries with two candidates or more), and only primary websites, we conducted the search process in seven steps:

1. Brute search: using an automated webscraper, we looked for all archived websites with URLs matching the following patterns: johnsmithforcongress.com, smithforcongress.com, johnsmithformassachusetts.com, smithformassachusetts.com, johnsmith2002.com, smith2002.com (using as example a candidate named John Smith, who ran in 2002 in Massachusetts). We only gathered websites that were captured during an election year. We found 2,650 potential websites.

2. Manual verification of brute search results: we hired a team of research assistants to verify all the websites found in step 1. We asked them to check three key elements on the website's main page: (i) that the website corresponds to the campaign for a House of Representatives race, (ii) that it contains the candidate's correct first and last name, and (iii) that it refers to the correct state. If any of these elements was missing, the website was categorized as "uncertain match." One of the paper's authors then checked all the uncertain matches manually. In total, 1,746 of the 2,650 websites found in step 1 were confirmed.

3. Manual search (round a): for all missing websites, we asked our research assistants to look manually for results on the Wayback Machine using the following keywords: john smith, john

smith massachusetts, smith massachusetts, john smith congress, smith congress, john smith house, smith house (using the same John Smith example as above). When in doubt, the team would categorize a website as "uncertain" and we would check it manually. We found an additional 1,719 websites through this step.

4. Manual search (round b): we hired another smaller team of more experienced research assistants to conduct the same procedure as step 3 for all websites that were still missing. We found an additional 274 websites through this step.
5. Manual search (round c): we ultimately searched for the last missing websites ourselves. We found an additional 8 websites through this step.
6. Automated verification: we ran a code on all websites collected through the steps above to verify whether their main page mentioned the candidate's first name, last name, and state. We identified 731 websites that lacked at least one of these elements.
7. Manual verification: an experienced research assistant manually verified all the 731 websites identified as potentially erroneous in the step above. After cross-verification by one of the paper's authors, we excluded 98 of these websites.

In total, we found and verified 3,649 House of Representatives primary election websites, out of 4,600 first- and second-ranked candidates running in a competitive primary between 2002 and 2016. After removing the websites for which the Wayback Machine captures were only taken *after* the primary election dates, we were left with 3,185 usable websites.

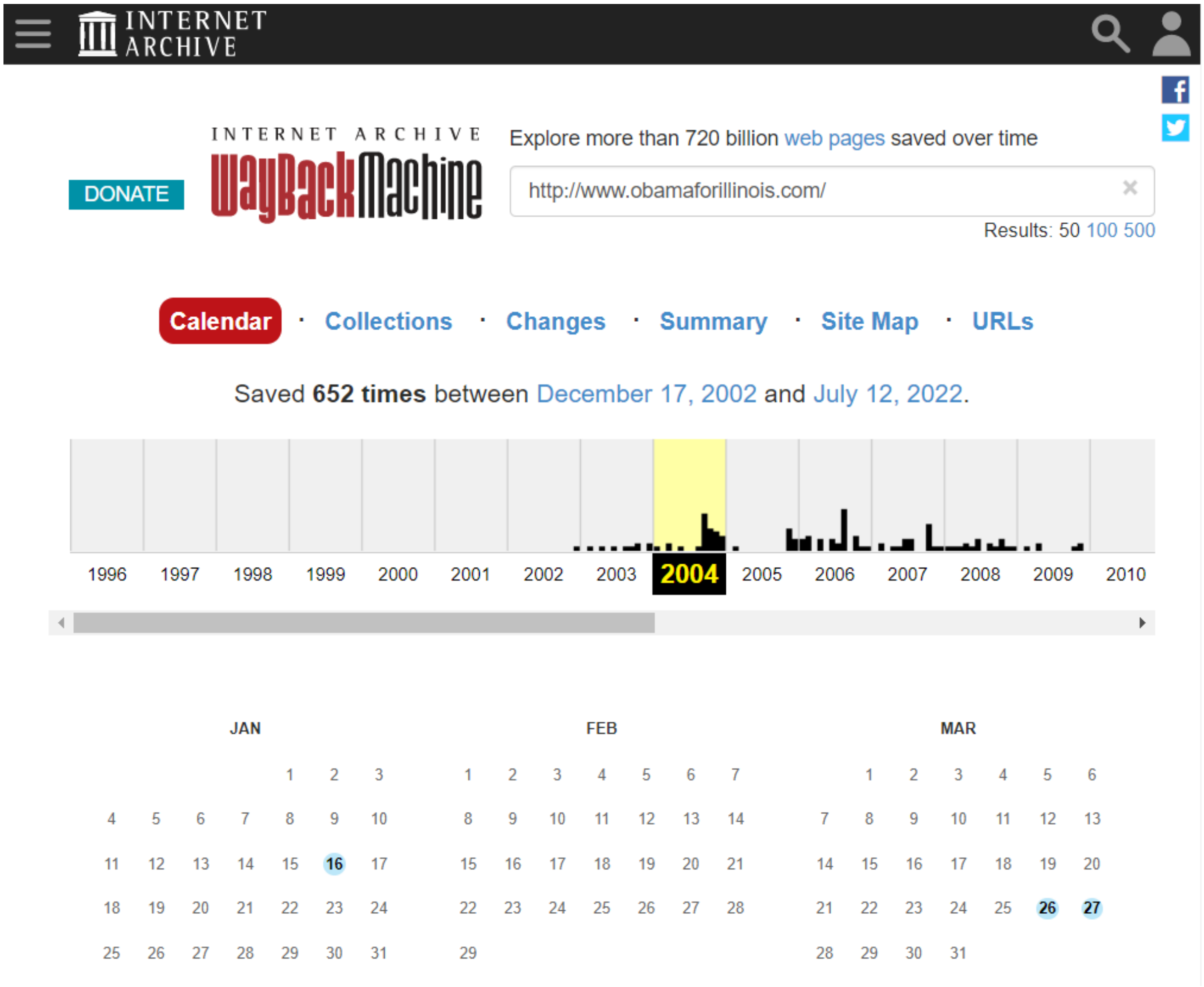
Collecting website content After retrieving the URLs pointing to the campaign websites, either from the Library of Congress (for the general election candidates) or from our own search on the Wayback Machine's interface (for the primary election candidates), we coded a Python scraper based on Selenium Web Driver to retrieve websites' content from the Wayback Machine. For each time capture of each campaign website available on the Wayback Machine, the scraper visited the main page, gathered all the textual content displayed on the page, identified hyperlinks, then visited all valid sub-pages, and gathered all the textual content displayed on these sub-pages. For computational reasons, we restrained the data collection to main pages and all sub-pages accessible within one click from main pages.

Web pages include HTML tags, which indicate titles, paragraphs, boxes, and so on. We coded the scraper so that the textual content of these different parts would remain separable once scraped and saved into CSV files.

We parallelized the scraper over 15 independent threads to make the process more time efficient. However, given the large number of pages visited, the Wayback Machine server sometimes reset the connection due to too many requests or failed to respond within the default allotted time (30 seconds). When this happened, the scraper was tasked to stop for a minute, then retry the procedure. If the scraper still failed to retrieve a time capture after the second attempt, it flagged an error in a separate CSV file and moved on to the next capture. After the scraper had attempted to retrieve all the time captures we had identified, we launched it again to try and correct the failed attempts. In the end, we only failed to retrieve 59 time captures out of 35,486.

Figure A.1: Example of website archiving

(a) Time captures



Explore more than 720 billion web pages saved over time

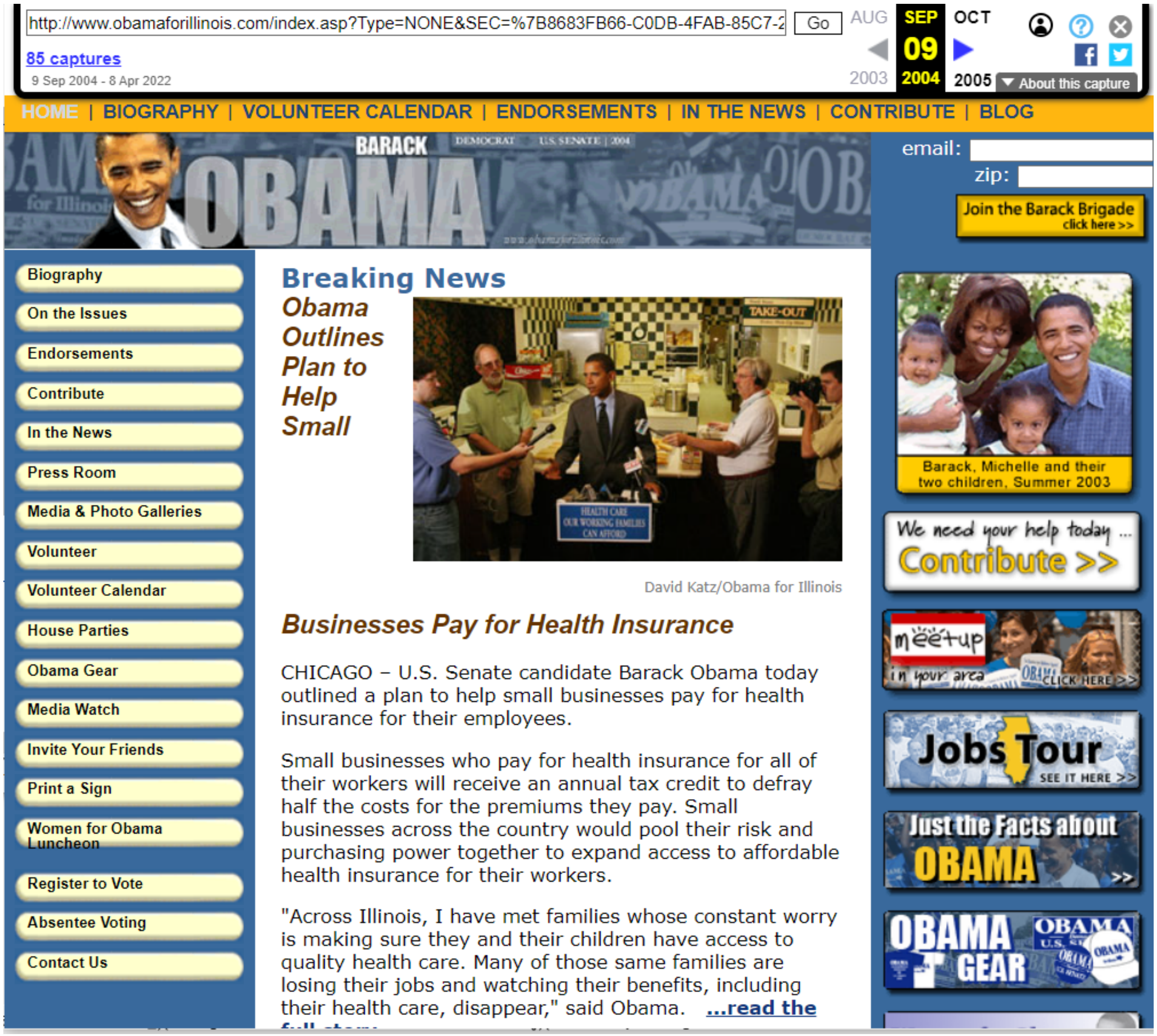
http://www.obamaforillinois.com/

Results: 50 100 500

Saved **652 times** between December 17, 2002 and July 12, 2022.

Figure A.1: Example of website archiving (cont.)

(b) Archived campaign webpage



Breaking News

Obama Outlines Plan to Help Small

David Katz/Obama for Illinois

Businesses Pay for Health Insurance

CHICAGO – U.S. Senate candidate Barack Obama today outlined a plan to help small businesses pay for health insurance for their employees.

Small businesses who pay for health insurance for all of their workers will receive an annual tax credit to defray half the costs for the premiums they pay. Small businesses across the country would pool their risk and purchasing power together to expand access to affordable health insurance for their workers.

"Across Illinois, I have met families whose constant worry is making sure they and their children have access to quality health care. Many of those same families are losing their jobs and watching their benefits, including their health care, disappear," said Obama. ...read the full story

Source: Wayback Machine

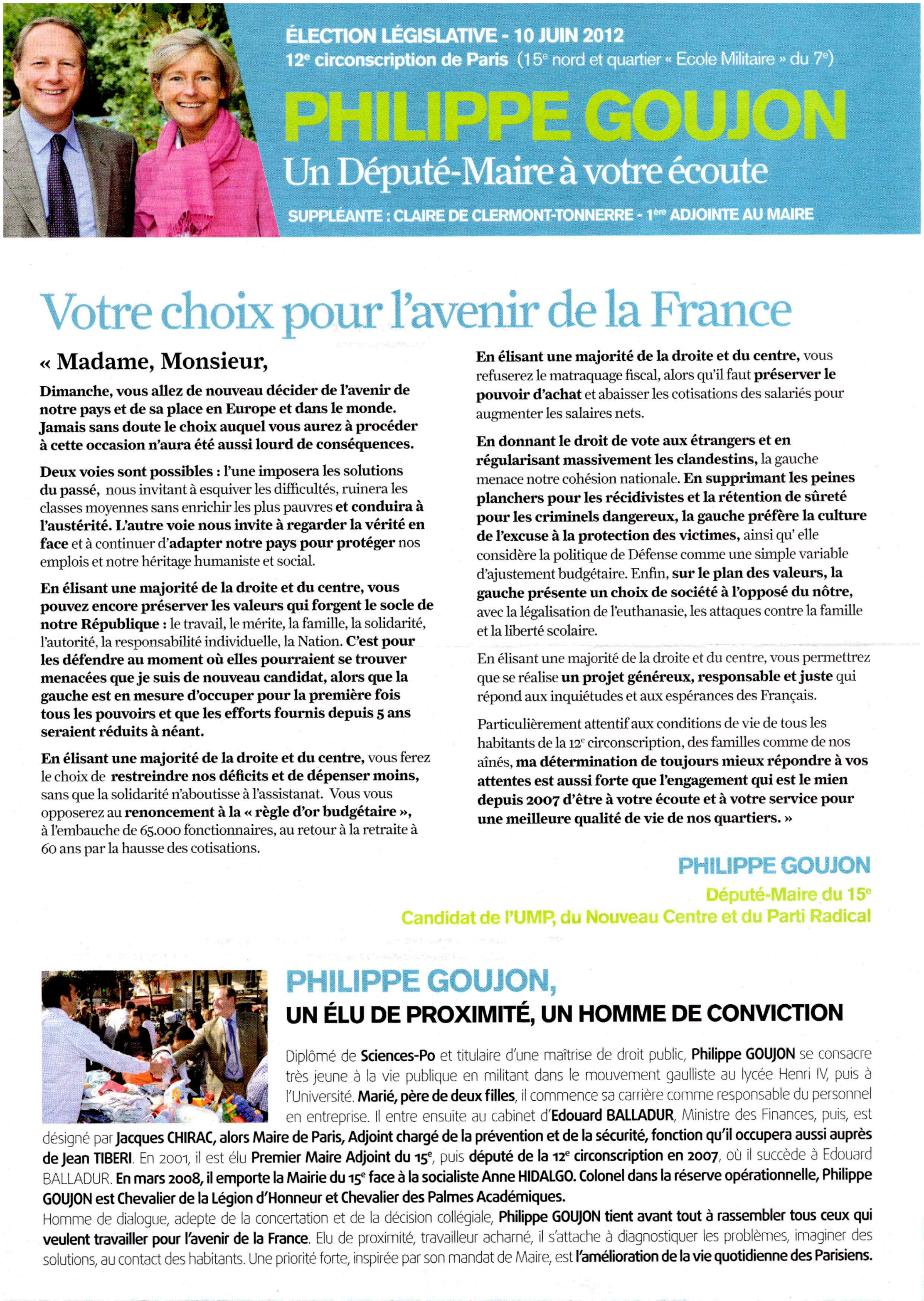
A.2 French candidate manifestos

In France, candidates are responsible for printing their electoral manifestos but the corresponding cost is refunded by the state if they gather at least 5% of the votes in the first round of the election (Electoral law, articles R39 and L216). Manifestos must have a maximum size of 210x297 millimeters, and a weight ranging between 60 and 80 grams per square meter (Electoral law, article R29). Furthermore, they cannot combine the three colors of the French flag (blue, white, and red), except if these are part of the party's emblem (Electoral law, article R27). The manifestos are mailed to registered voters up to four days before the election (for the first round), and three days before the second round when there is a runoff (Electoral law, articles R34 and R38).

TV shows and TV debates remain the prominent medium for candidates campaigning at the national level, such as presidential candidates and party leaders advertising their party's national platform. However, the information provided on television is unlikely to teach voters much about the individual candidates running in their local or parliamentary constituency. Candidate manifestos are an important vector for these candidates to tailor their message to the voters in their district. In a survey published before the 2017 election, 24% of respondents declared that manifestos were one of the three most important ways for them to get information about the candidates (OpinionWay, 2017). By comparison, television was mentioned by 64% of them, online media by 26%, printed newspapers by 18%, and radio by 15%.

Figure A.2: Example of candidate manifesto

(a) First page



ÉLECTION LÉGISLATIVE - 10 JUIN 2012
12e circonscription de Paris (15e nord et quartier « Ecole Militaire » du 7e)

PHILIPPE GOUJON

Un Député-Maire à votre écoute

SUPPLÉANTE : CLAIRE DE CLERMONT-TONNERRE - 1ère ADJOINTE AU MAIRE

Votre choix pour l'avenir de la France

« Madame, Monsieur,

Dimanche, vous allez de nouveau décider de l'avenir de notre pays et de sa place en Europe et dans le monde. Jamais sans doute le choix auquel vous aurez à procéder à cette occasion n'aura été aussi lourd de conséquences.

Deux voies sont possibles : l'une imposera les solutions du passé, nous invitant à esquiver les difficultés, ruinera les classes moyennes sans enrichir les plus pauvres **et conduira à l'austérité. L'autre voie nous invite à regarder la vérité en face** et à continuer d'**adapter notre pays pour protéger** nos emplois et notre héritage humaniste et social.

En élisant une majorité de la droite et du centre, vous pouvez encore préserver les valeurs qui forgent le socle de notre République : le travail, le mérite, la famille, la solidarité, l'autorité, la responsabilité individuelle, la Nation. **C'est pour les défendre au moment où elles pourraient se trouver menacées que je suis de nouveau candidat, alors que la gauche est en mesure d'occuper pour la première fois tous les pouvoirs et que les efforts fournis depuis 5 ans seraient réduits à néant.**

En élisant une majorité de la droite et du centre, vous ferez le choix de **restreindre nos déficits et de dépenser moins,** sans que la solidarité n'aboutisse à l'assistanat. Vous vous opposerez au **renoncement à la « règle d'or budgétaire »,** à l'embauche de 65.000 fonctionnaires, au retour à la retraite à 60 ans par la hausse des cotisations.

En élisant une majorité de la droite et du centre, vous refuserez le matraquage fiscal, alors qu'il faut **préserver le pouvoir d'achat** et abaisser les cotisations des salariés pour augmenter les salaires nets.

En donnant le droit de vote aux étrangers et en régularisant massivement les clandestins, la gauche menace notre cohésion nationale. **En supprimant les peines planchers pour les récidivistes et la rétention de sûreté pour les criminels dangereux, la gauche préfère la culture de l'excuse à la protection des victimes,** ainsi qu' elle considère la politique de Défense comme une simple variable d'ajustement budgétaire. Enfin, **sur le plan des valeurs, la gauche présente un choix de société à l'opposé du nôtre,** avec la légalisation de l'euthanasie, les attaques contre la famille et la liberté scolaire.

En élisant une majorité de la droite et du centre, vous permettrez que se réalise **un projet généreux, responsable et juste** qui répond aux inquiétudes et aux espérances des Français.

Particulièrement attentif aux conditions de vie de tous les habitants de la 12e circonscription, des familles comme de nos aînés, **ma détermination de toujours mieux répondre à vos attentes est aussi forte que l'engagement qui est le mien depuis 2007 d'être à votre écoute et à votre service pour une meilleure qualité de vie de nos quartiers. »**

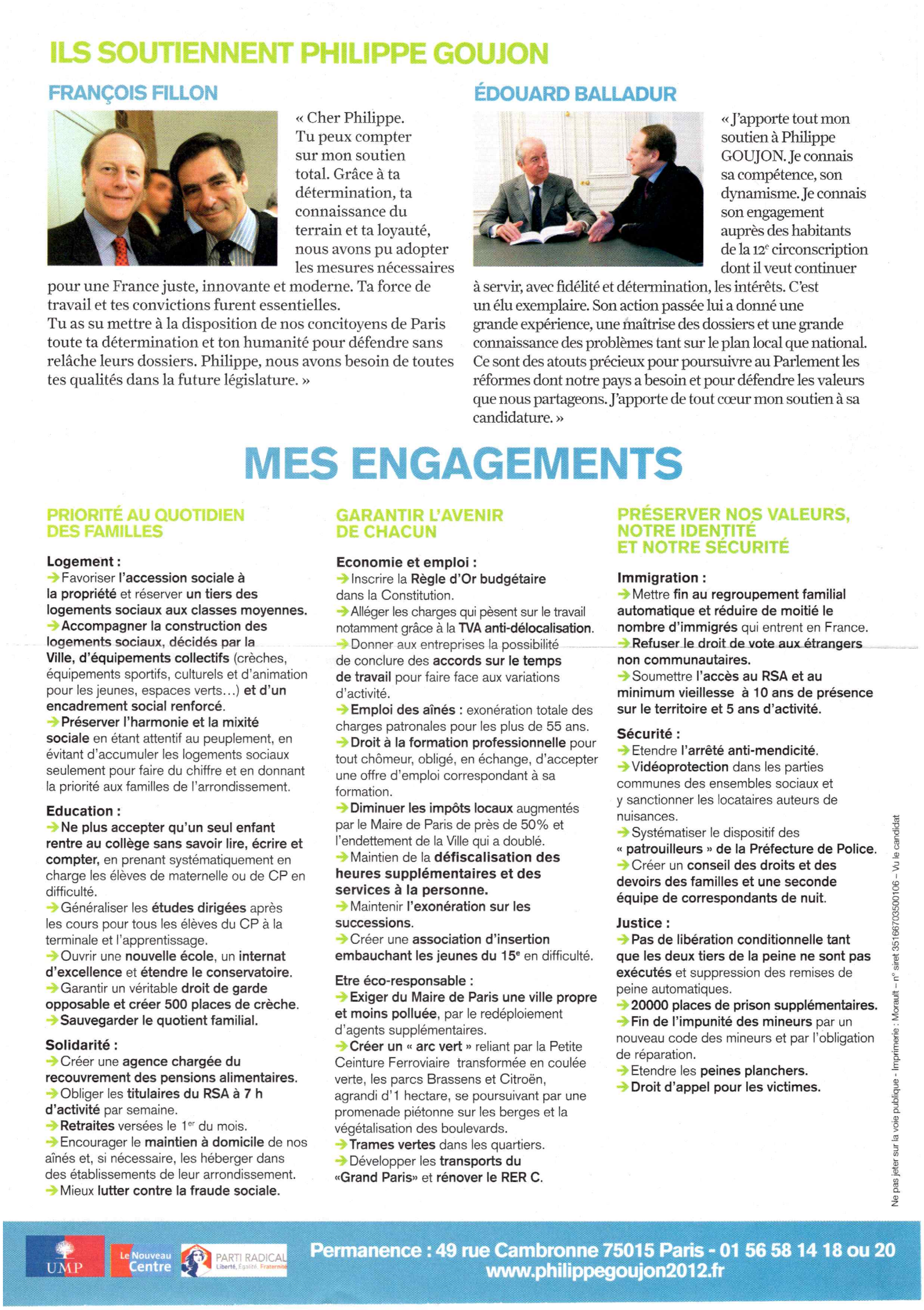
PHILIPPE GOUJON
Député-Maire du 15e
Candidat de l'UMP, du Nouveau Centre et du Parti Radical

PHILIPPE GOUJON,
UN ÉLU DE PROXIMITÉ, UN HOMME DE CONVICTION

Diplômé de **Sciences-Po** et titulaire d'une maîtrise de droit public, **Philippe GOUJON** se consacre très jeune à la vie publique en militant dans le mouvement gaulliste au lycée Henri IV, puis à l'Université. **Marié, père de deux filles**, il commence sa carrière comme responsable du personnel en entreprise. Il entre ensuite au cabinet d'**Edouard BALLADUR**, Ministre des Finances, puis, est désigné par **Jacques CHIRAC, alors Maire de Paris, Adjoint chargé de la prévention et de la sécurité, fonction qu'il occupera aussi auprès de Jean TIBERI**. En 2001, il est élu **Premier Maire Adjoint du 15e**, puis **député de la 12e circonscription en 2007**, où il succède à Edouard BALLADUR. **En mars 2008, il emporte la Mairie du 15e face à la socialiste Anne HIDALGO. Colonel dans la réserve opérationnelle, Philippe GOUJON est Chevalier de la Légion d'Honneur et Chevalier des Palmes Académiques.**
Homme de dialogue, adepte de la concertation et de la décision collégiale, **Philippe GOUJON tient avant tout à rassembler tous ceux qui veulent travailler pour l'avenir de la France**. Elu de proximité, travailleur acharné, il s'attache à diagnostiquer les problèmes, imaginer des solutions, au contact des habitants. Une priorité forte, inspirée par son mandat de Maire, est **l'amélioration de la vie quotidienne des Parisiens.**

Figure A.2: Example of candidate manifesto (cont.)

(b) Second page



ILS SOUTIENNENT PHILIPPE GOUJON

FRANÇOIS FILLON

« Cher Philippe. Tu peux compter sur mon soutien total. Grâce à ta détermination, ta connaissance du terrain et ta loyauté, nous avons pu adopter les mesures nécessaires pour une France juste, innovante et moderne. Ta force de travail et tes convictions furent essentielles.
Tu as su mettre à la disposition de nos concitoyens de Paris toute ta détermination et ton humanité pour défendre sans relâche leurs dossiers. Philippe, nous avons besoin de toutes tes qualités dans la future législature. »

ÉDOUARD BALLADUR

« J'apporte tout mon soutien à Philippe GOUJON. Je connais sa compétence, son dynamisme. Je connais son engagement auprès des habitants de la 12ᵉ circonscription dont il veut continuer à servir, avec fidélité et détermination, les intérêts. C'est un élu exemplaire. Son action passée lui a donné une grande expérience, une maîtrise des dossiers et une grande connaissance des problèmes tant sur le plan local que national. Ce sont des atouts précieux pour poursuivre au Parlement les réformes dont notre pays a besoin et pour défendre les valeurs que nous partageons. J'apporte de tout cœur mon soutien à sa candidature. »

MES ENGAGEMENTS

PRIORITÉ AU QUOTIDIEN DES FAMILLES

Logement :

→ Favoriser **l'accession sociale à la propriété** et réserver **un tiers des logements sociaux aux classes moyennes.**

→ **Accompagner la construction des logements sociaux, décidés par la Ville, d'équipements collectifs** (crèches, équipements sportifs, culturels et d'animation pour les jeunes, espaces verts…) **et d'un encadrement social renforcé.**

→ **Préserver l'harmonie et la mixité sociale** en étant attentif au peuplement, en évitant d'accumuler les logements sociaux seulement pour faire du chiffre et en donnant la priorité aux familles de l'arrondissement.

Education :

→ **Ne plus accepter qu'un seul enfant rentre au collège sans savoir lire, écrire et compter,** en prenant systématiquement en charge les élèves de maternelle ou de CP en difficulté.

→ Généraliser les **études dirigées** après les cours pour tous les élèves du CP à la terminale et l'apprentissage.

→ Ouvrir une **nouvelle école**, un **internat d'excellence** et **étendre le conservatoire.**

→ Garantir un véritable **droit de garde opposable et créer 500 places de crèche.**

→ **Sauvegarder le quotient familial.**

Solidarité :

→ Créer une **agence chargée du recouvrement des pensions alimentaires.**

→ Obliger les **titulaires du RSA à 7 h d'activité** par semaine.

→ **Retraites** versées le 1ᵉʳ du mois.

→ Encourager le **maintien à domicile** de nos aînés et, si nécessaire, les héberger dans des établissements de leur arrondissement.

→ Mieux **lutter contre la fraude sociale.**

GARANTIR L'AVENIR DE CHACUN

Economie et emploi :

→ Inscrire la **Règle d'Or budgétaire** dans la Constitution.

→ Alléger les charges qui pèsent sur le travail notamment grâce à la **TVA anti-délocalisation.**

→ Donner aux entreprises la possibilité de conclure des **accords sur le temps de travail** pour faire face aux variations d'activité.

→ **Emploi des aînés :** exonération totale des charges patronales pour les plus de 55 ans.

→ **Droit à la formation professionnelle** pour tout chômeur, obligé, en échange, d'accepter une offre d'emploi correspondant à sa formation.

→ **Diminuer les impôts locaux** augmentés par le Maire de Paris de près de 50% et l'endettement de la Ville qui a doublé.

→ Maintien de la **défiscalisation des heures supplémentaires et des services à la personne.**

→ Maintenir **l'exonération sur les successions.**

→ Créer une **association d'insertion embauchant les jeunes du 15ᵉ** en difficulté.

Etre éco-responsable :

→ **Exiger du Maire de Paris une ville propre et moins polluée**, par le redéploiement d'agents supplémentaires.

→ **Créer un « arc vert »** reliant par la Petite Ceinture Ferroviaire transformée en coulée verte, les parcs Brassens et Citroën, agrandi d'1 hectare, se poursuivant par une promenade piétonne sur les berges et la végétalisation des boulevards.

→ **Trames vertes** dans les quartiers.

→ Développer les **transports du «Grand Paris»** et **rénover le RER C.**

PRÉSERVER NOS VALEURS, NOTRE IDENTITÉ ET NOTRE SÉCURITÉ

Immigration :

→ Mettre **fin au regroupement familial automatique et réduire de moitié le nombre d'immigrés** qui entrent en France.

→ **Refuser le droit de vote aux étrangers non communautaires.**

→ Soumettre **l'accès au RSA et au minimum vieillesse à 10 ans de présence sur le territoire et 5 ans d'activité.**

Sécurité :

→ Etendre **l'arrêté anti-mendicité.**

→ **Vidéoprotection** dans les parties communes des ensembles sociaux et y sanctionner les locataires auteurs de nuisances.

→ Systématiser le dispositif des **« patrouilleurs » de la Préfecture de Police.**

→ Créer un **conseil des droits et des devoirs des familles et une seconde équipe de correspondants de nuit.**

Justice :

→ **Pas de libération conditionnelle tant que les deux tiers de la peine ne sont pas exécutés** et suppression des remises de peine automatiques.

→ **20000 places de prison supplémentaires.**

→ **Fin de l'impunité des mineurs** par un nouveau code des mineurs et par l'obligation de réparation.

→ Etendre les **peines planchers.**

→ **Droit d'appel pour les victimes.**

Ne pas jeter sur la voie publique - Imprimerie : Morault – n° siret 35166703500106 – Vu le candidat

Permanence : 49 rue Cambronne 75015 Paris - 01 56 58 14 18 ou 20
www.philippegoujon2012.fr

Source: Paris Municipal Archives.

Data collection Manifestos issued for the parliamentary elections held between 1958 and 1993 were systematically collected and digitized by the CEVIPOF and the Sciences Po Library for the Archelec project (Gaultier-Voituriez, 2016). We use the dataset assembled by Le Pennec (2023), which links the content of each manifesto to the electoral results' database using fuzzy string matching and hand-coding of candidates' names. It contains 24,431 first-round manifestos and 6,885 second-round manifestos.

Manifestos issued for the 1997 parliamentary elections were collected from the National Archives by Cagé, Le Pennec and Mougin (2023). The authors scanned and digitized the paper documents using optical character recognition, and they linked the obtained machine-readable content of each manifesto to the electoral results database using fuzzy string matching on candidates' names. These data contain 5,356 first-round manifestos and 1,039 second-round manifestos.

A subset of the manifestos issued for the 2017 parliamentary elections were published online in PDF version by the Ministry of the Interior shortly before the election, and scraped and turned into machine-readable text by the non-profit organization *Regards Citoyens*.[45] These manifestos were linked to the electoral results database by Le Pennec (2023), using fuzzy string matching on candidates' names. These data contain 4,981 first-round manifestos and 702 second-round manifestos.

Manifestos issued for the 2021 local elections were also collected by *Regards Citoyens* and we linked them to the electoral database using candidates' ballot registration number. When possible, we extracted the textual content directly from the PDF documents using the PyPDF2 and Tika libraries in Python (about 74% of manifestos). The remaining PDF documents were turned into machine-readable text using the Tesseract OCR-engine. These data contain 4,451 first-round manifestos and 27 second-round manifestos.

We collected the 2022 manifestos made available online in PDF version by the Ministry of the Interior using a web scraper that we coded based on the Python implementation of the Selenium Web Driver and BeautifulSoup. We linked the manifestos to the electoral database using candidates' ballot registration number available online along with the PDF documents. When possible, we extracted the textual content directly from the PDF documents using the PyPDF2 and Tika libraries in Python (about 85% of manifestos). The remaining PDF documents were turned into machine-readable text using the Tesseract OCR-engine. These data contain 4,832 first-round manifestos and 1,020 second-round manifestos. Candidates to the 2022 elections also had the possibility to submit an alternative version of their manifesto meant to be "easy to read and understand" to be published online along with their main manifesto. We collected 2,363 such manifestos for the first round and 636 for the second round.

Finally, we made a systematic effort at hand-collecting additional manifestos in Ile-de-France, the

[45] See: https://www.regardscitoyens.org.

most populated region that includes the city of Paris and seven other *départements*: Essonne, Hauts-de-Seine, Seine-Saint-Denis, Seine-et-Marne, Val-de-Marne, Val-d'Oise, and Yvelines. In each of these *départements*, we contacted all local administrations likely to have collected candidate manifestos: *Préfectures*, Departmental councils, Departmental archives, the town hall of each *département*'s capital city, the municipal archives of each *département*'s capital city, and the public multimedia library of each *département*'s capital city. We also contacted the local headquarters of the Socialist Party, the Communist Party, and the Republican Party (formerly UMP and RPR). Out of the 98 organizations that we contacted, 26 responded that they had paper versions of candidate manifestos for the elections that we targeted in priority: the 2002, 2007, and 2012 parliamentary elections. We visited each of these places and digitized all the manifestos available for these three elections as well as all the manifestos available for local elections since 1979 and for the parliamentary elections between 1958 and 1997 that were missing from other data sources. We used the Tesseract OCR-engine to turn the PDF documents into machine-readable text and linked each manifesto to electoral results at the candidate level with fuzzy string matching. This data collection added 2,733 first-round manifestos and 632 second-round manifestos to our dataset.

Table A.1: U.S. sampling frame

	Primary elections			General elections		
Year	Races	Candidates	Websites	Races	Candidates	Websites
2002	158	465	224	322	925	565
2004	198	519	302	365	1,000	687
2006	200	608	304	379	1,000	715
2008	225	629	357	391	1,052	772
2010	362	1,123	574	412	1,242	860
2012	328	916	527	403	1,104	780
2014	262	694	442	378	979	704
2016	289	849	455	386	1,023	709

Notes: For each election at the U.S. House of Representatives, we indicate the number of races for which we have collected at least one website, the number of candidates in these races, and the number of candidates for which a website is available, for the primary and general election separately.

Table A.2: France sampling frame

(a) Parliamentary elections

	First round			Second round		
Year	Races	Candidates	Manifestos	Races	Candidates	Manifestos
1958	361	2,060	1,947	277	871	803
1962	465	2,171	1,699	351	872	535
1967	461	2,135	2,052	385	846	822
1968	465	2,246	2,220	300	647	642
1973	473	3,092	2,920	424	946	919
1978	469	4,140	3,950	410	813	812
1981	474	2,557	2,403	318	627	626
1988	502	2,469	2,374	388	775	764
1993	554	5,130	4,866	482	962	962
1997	543	6,049	5,421	500	1,066	1,041
2002	40	640	201	31	63	55
2007	55	758	395	42	84	80
2012	31	400	350	42	82	82
2017	565	7,682	4,969	455	911	701
2022	563	6,121	4,809	549	1,102	1,016

(b) Local elections

	First round			Second round		
Year	Races	Candidates	Manifestos	Races	Candidates	Manifestos
1979	60	289	275	29	57	51
1982	58	298	229	34	68	59
1985	61	442	328	36	75	67
1988	33	202	156	19	30	27
2001	43	286	260	39	77	71
2004	1	8	8	2	5	2
2008	41	216	145	24	48	36
2011	47	322	249	45	86	73
2015	9	50	50	10	20	20
2021	1,782	7,093	4,331	15	31	24

Notes: For each French parliamentary and local election, we indicate the number of races for which we have collected at least one manifesto, the number of candidates in these races, and the number of candidates for which a manifesto is available, for the first and second round separately.

A.3 Voter characteristics

We complement our database of candidate websites and manifestos with information on voter characteristics.

In the U.S., we retrieved sociodemographic information from the data made publicly available by the census.[46] Specifically, we collected data at the congressional level from the 2010 American Community Survey and subsequent forecasts on the population's average age, income, education, citizenship status, and employment rate from 2010 to 2016. We also obtained congressional districts' population density from the CityLab.[47]

In France, we retrieved sociodemographic data from the French national statistics agency (INSEE).[48] Specifically, we collected data at the municipality level on the population's average age, income, education, citizenship status, employment rate, and population density. Some of these variables (e.g., citizenship status) are only available on the INSEE website post 2007, hence we focus on elections occurring after this date when using census data in France. When the census is not available for a given election year post 2007, we use the closest available year (e.g., 2018 for 2022). We then aggregated these outcomes at the district level, using municipalities' population as weights.

We completed these sociodemographic data with presidential vote shares from MIT Election Data and Science Lab (2017*b*) for the 2008, 2012, and 2014 U.S. elections, and from the Ministry of the Interior[49] for the 2002, 2007, 2012, and 2017 French elections. We use these results as a proxy for districts' political orientation.

[46] See: https://data.census.gov/table?q=All+Congressional+Districts+within+United+States.

[47] See: https://github.com/theatlantic/citylab-data/tree/master/citylab-congress.

[48] See: https://www.insee.fr/fr/statistiques.

[49] See: https://www.data.gouv.fr/fr/pages/donnees-des-elections/.

B Additional Details on the Text Analysis

B.1 Text pre-processing

In the U.S., we pre-process the websites' content by removing all URLs, numbers, and special characters except for basic punctuation (?!'-,;). Additionally, we discard the parts of websites delineated by HTML tags containing less than 10 words (e.g., navigation bar, headers). In France, we also pre-process the manifestos by removing special characters except for basic punctuation and numbers. We replace accented letters with their unaccented equivalent.

In both cases, we transform the text into lower-case and tokenize documents at the single-word level. For the vector representations, we also stem words using NLTK's SnowballStemmer in Python in order to improve the training efficiency.

B.2 Ideological score

Vocabulary Prior to calculating word scores for the U.S sample, we exclude words used by fewer than 0.5% and more than 80% of all Democratic and Republican primary election candidates, in a given election year. This leaves us with an average vocabulary of 11,400 words per election year.

Similarly, in the French sample, we exclude words used by fewer than 0.5% and more than 80% of all left-wing and right-wing first-round candidates, in a given election year. This leaves us with an average vocabulary of 6,100 words per election year.

Score normalization After calculating a document j's "raw" score S_j, we implement the normalization proposed by Martin and Vanberg (2007), so that the final ideological score is defined as:

$$Score_j = \frac{S_j}{S^R},$$

where $S^R = \sum_w p_w^R \cdot s_w$ is the estimated ideological score of an average right-wing document, p_w^R is the average frequency of word w among right-wing documents, and s_w is the word score of word w as defined in Section 3.1. This normalization ensures that the original distance between right-wing and left-wing manifestos or websites is preserved in the estimated score dispersion. Hence, the final partisan score is not bounded between -1 and 1. Instead, a score of 1 corresponds to the score of a document that is representative of the average ideology on the right side, while a score of -1 corresponds to the score of a document that is representative of the average ideology on the left side.

Validation Tables B.1 and B.2 show the twenty words with the highest (most right) and lowest (most left) score for the U.S. and French elections. Since we calculate word scores in each election year separately, each score shown in these tables corresponds to the word's average score across all election years.

Table B.1: U.S. lowest and highest ideological word scores

(a) Left-wing words

Word	Ideology score
polluters	-0.98
wealthiest	-0.91
minorities	-0.79
trades	-0.73
disproportionately	-0.72
howard	-0.71
equality	-0.70
richest	-0.69
renewable	-0.67
longest	-0.66
divisive	-0.66
universities	-0.65
transit	-0.64
electrical	-0.64
loophole	-0.63
solar	-0.63
counseling	-0.63
pollution	-0.63
discrimination	-0.63
gay	-0.62

(b) Right-wing words

Word	Ideology score
unborn	0.98
sanctity	0.93
liberals	0.93
aliens	0.92
conservatism	0.89
beef	0.87
bureaucrats	0.83
amnesty	0.83
pray	0.75
babies	0.72
abortions	0.72
libertarian	0.72
bible	0.71
intrusion	0.71
upholding	0.70
sportsmen	0.70
contrary	0.69
arctic	0.68
principled	0.68
ross	0.65

Notes: We list the 20 words with the lowest (Panel a) and highest (Panel b) ideological word score over the sample period (averaging, for each word, the scores in each election year). We rank words from all tokens used by at least 0.5 % and at most 80 % of primary election websites by Democratic or Republican candidates, in every election year.

Table B.2: France lowest and highest ideological word scores

(a) Left-wing words

Word	Translation	Ideology score
antisociales	antisocial	-1.00
feministe	feminist	-0.99
ogm	GMO	-0.98
reduisent	reduce	-0.97
licencient	lay off	-0.96
chevenement	(see note)	-0.95
pesticides	pesticides	-0.94
pesera	will weigh	-0.89
laiques	secular	-0.88
considerent	consider	-0.86
partages	sharing	-0.83
alternatifs	alternative	-0.82
ultra	ultra	-0.82
laisseront	will let	-0.80
opprimes	oppressed	-0.79
verses	deposited	-0.78
agences	agencies	-0.76
tales	hit	-0.76
exploites	exploited	-0.76
situa	situation	-0.75

(b) Right-wing words

Word	Translation	Ideology score
socialocommuniste	social-communist	0.97
terroirs	land	0.91
gendarmes	military police	0.79
sejour	stay	0.77
brigade	brigade	0.76
terrorisme	terrorism	0.76
perils	dangers	0.72
postal	postal	0.70
clandestine	clandestine	0.69
exportations	exports	0.65
titres	headlines	0.64
optique	optic	0.63
sauvegardant	safeguarding	0.62
automobilistes	car drivers	0.60
vehicule	car	0.59
assurons	ensure	0.59
formalites	formalities	0.59
irresponsabilite	irresponsibility	0.57
patriotisme	patriotism	0.56
totalitaire	totalitarian	0.55

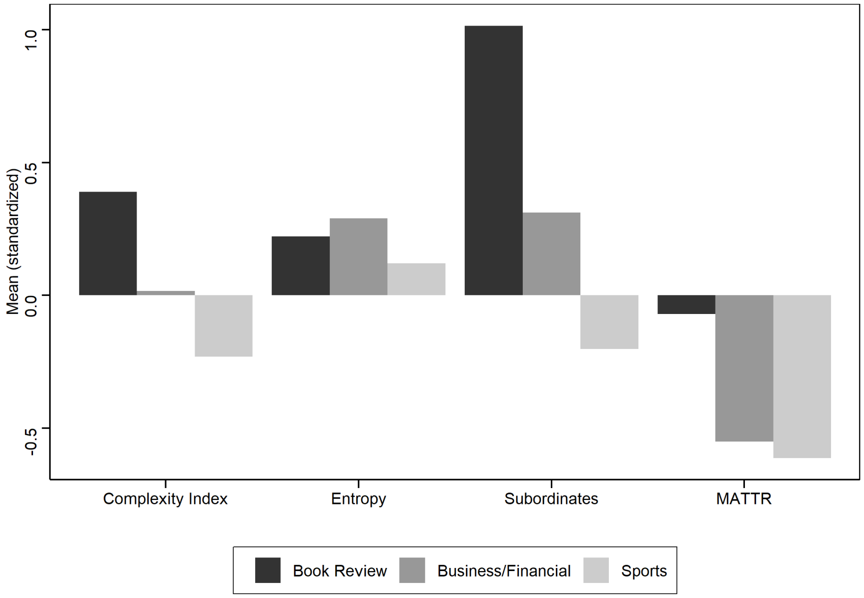
Notes: We list the 20 words with the lowest (Panel a) and highest (Panel b) ideological word score over the sample period (averaging, for each word, the scores in each election year). We rank words from all tokens used by at least 0.5 % and at most 80 % of first-round manifestos by left-wing or right-wing candidates, in every election year. "Chevenement" refers to the 1998 Chevènement law that aim to facilitate migrant families' reunification.

B.3 Complexity

Validation Candidates at the 2022 French parliamentary election had the possibility to publish an alternative version of their manifesto meant to be "easy to read and understand" (*facile à lire et à comprendre* or *FALC*) along with their original manifesto, and 2,989 did so. We use these manifestos to benchmark our complexity metrics. Figure B.2 shows that our complexity index deems the regular manifestos to be more complex than their FALC version, as expected. In particular, the FALC manifestos use substantively simpler words (lower Entropy) and less diverse words (lower MATTR). However, they use more structurally complex sentences (higher Subordinates). This could be explained by the fact that conveying the same ideas with simpler words requires longer sentences.

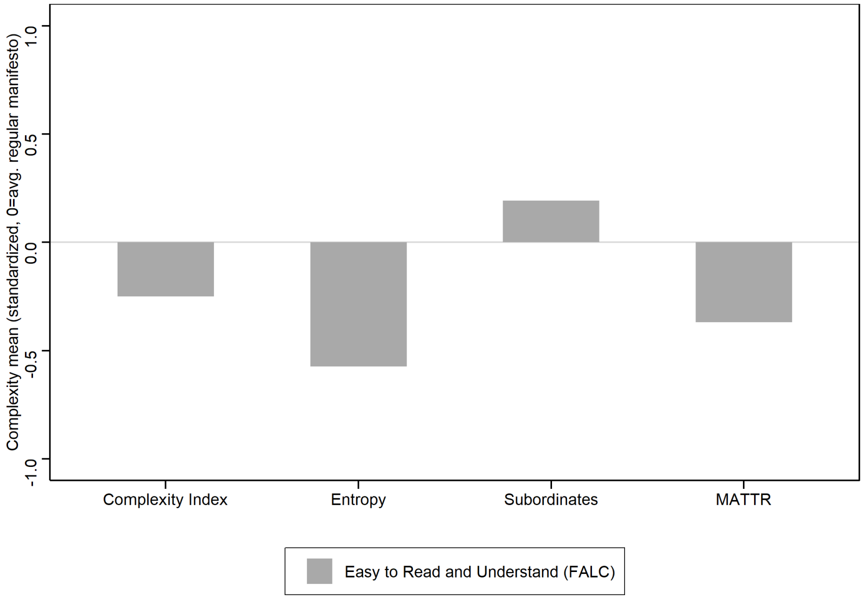
In the U.S., we benchmark websites' complexity against articles from different sections of the *New York Times*. We use Factiva to download 1,000 articles published in the *New York Times* in August 2022, measure each article complexity, and aggregate complexity scores by section. We standardize the different complexity components using the mean and standard deviation from the candidates' websites, where the complexity of a website is defined as the average complexity across a candidate's website captures taken before the primary election or between the primary and the general election. Results are shown in Figure B.1. The average complexity of candidates' websites is equivalent to the articles published in the Business/Financial section of the *New York Times*, lower than the Book Review section and higher than the Sports section. The difference between the Sports and Book Review sections is about 0.6 standard deviation. Looking at the different complexity metrics that enter in the index, we find that the average website uses less complex words (Entropy) than all sections, sentence structures that are more complex than the Sports section but less complex than the Business/Financial and Book Review sections (Subordinates), and addresses more diverse subjects than all sections.

Figure B.1: U.S. websites complexity compared to the New York Times



Notes: The horizontal line at 0 represents, for each complexity measure, the average complexity of a candidate website among the 5,792 general election candidates for which we have found a website. Each bar represents the standardized complexity of an article in a given section, among 1,000 articles published in August 2022 in the *New York Times*, relative to the average candidate website.

Figure B.2: Manifestos' complexity comparison



Notes: The horizontal line at 0 represents, for each complexity measure, the average complexity of a regular manifesto, among the 2,989 manifestos issued in 2022 with a FALC equivalent. Each bar represents the standardized average complexity of a FALC manifesto, relative to the average complexity of a regular manifesto.

B.4 Topic distribution

Policy topics The Manifesto Project classifies sentences of national manifestos into 84 narrow topics (e.g., "Military: Positive," "Education Expansion," "Agriculture and Farmers: Negative," etc.). We group these subtopics under 31 larger topics (e.g., "Military: Positive" is grouped under "Military," "Education Expansion" under "Education," "Agriculture and Farmers: Negative" under "Agriculture and Farmers," etc.). There is a small number of narrow topics, covering less than 0.5% of the sentences in the U.S. manifestos, that are not obvious subtopics of a larger topic (e.g., "Marxist Analysis"). We assign them to the closest large topic (e.g., "Marxist Analysis" is grouped under "Other").

The Agenda Project classifies sentences of national manifestos into 27 topics (e.g., "Work and Employment," "Social groups," etc.) which we use as given.

The final list of topics for each country is shown in Table B.3.

Method To quantify the relative importance of different topics in candidates' communication, we implement a supervised machine learning model trained on the manifestos issued by national parties in the U.S. and in France. First, we transform party manifestos into vectors using a TF-IDF vectorizer. Then, we feed the TF-IDF vectors into an SVM classifier to predict each topic's likelihood of being addressed in a given training sentence. Note that SVMs do not directly provide probability distributions. Estimating these probabilities requires an additional step called Platt scaling, which is transparent to the user thanks to the sklearn Python library.[50]

We explored several options to select the best performing model for this classification task, namely a linear model, a logistic regression, a random forest, and several gradient boosted random forest classifiers. Performance was assessed using the average accuracy over five-fold cross validations. The SVM classifier yielded the best average accuracy: 56% in the U.S. and 51% in France. These numbers represent substantial improvements over a random allocation of 30 topics across documents, which would yield an average accuracy of 3%. After selecting the SVM classifier, we further performed a grid search to optimize the model's hyperparameters (kernel, gamma function, and regularisation parameter).

Koh, Boey and Béchara (2021) show that using deep learning models (Convolutional Neural Networks) in conjunction with state-of-the-art language models (BERT) only provides a marginal improvement (a difference of 0.2 percentage point in F1-score) compared to a TFIDF-SVM pipe. This marginal improvement comes at a high computational and time cost, hence we decided to keep using the TFIDF-SVM pipe.

[50] See `https://scikit-learn.org/stable/modules/svm.html`.

Most predictive words Table B.3 lists the most predictive words, as given by our trained SVM model, associated with each topic in the French and U.S. national manifestos.

Table B.3: Most predictive words associated with topics

(a) U.S.

Topic	Most predictive stemmed words
Agriculture and Farmers	agricultur, farm, farmer, food, suffici, rural, biofuel, farmwork, crop, forest
Centralisation	state, feder, overfeder, local, washington, territori, selfsuffici, resourc, string, island
Civic Mindedness	communiti, togeth, neighborhood, civic, serv, allegi, where, everyon, uniti, trust
Constitutionalism	constitut, amend, document, recess, unconstitut, judiciari, appoint, oval, firearm, senat
Culture	art, tourism, fish, artist, endow, recreat, cultur, heritag, rejuven, hunt
Democracy	democraci, vote, voter, democrat, liberti, elect, independ, peopl, ballot, congress
Economic Planning	econom, spend, economi, prosper, growth, debt, fiscal, deficit, cap, budget
Education	educ, school, colleg, student, teacher, period, basic, learn, read, afterschool
Equality	discrimin, disabl, equal, women, racial, color, access, poverti, inequ, racism
Foreign Special Relationships	israel, cuba, iraqi, regim, coalit, latin, dictatorship, canada, relationship, iranian
Free Market Economy	regul, properti, weaken, enterpris, permit, privat, sector, radic, minimum, market
Freedom and Human Rights	privaci, right, tortur, freedom, humanitarian, digniti, human, control, free, journalist
Governmental and Administrative Efficiency	effici, backlog, simplifi, downsiz, better, simpler, depart, agenc, govern, wast
Incentives	tax, busi, entrepreneur, incent, entrepreneurship, small, taxat, paperwork, key, lower
Internationalism	global, diplomaci, partnership, nato, mexico, intern, un, foreign, hivaid, europ
Labour Groups	job, worker, union, workforc, work, workplac, wage, labor, workers, unemploy
Law and Order	crime, intellig, sentenc, crimin, law, terrorist, safer, penalti, redempt, prosecut
Market Regulation	financi, poorer, antitrust, consum, top, crack, bailout, street, loan, wall
Military	militari, defens, troop, secur, forc, nuclear, defend, isi, nonprolifer, guard
Multiculturalism	tribal, indian, nativ, immigr, indigen, tribe, alien, nationton, cultur, divers
National Way of Life	legal, bounti, charact, histori, earth, valu, visa, asylum, idea, soul
Peace	peac, palestinian, rivalri, sudan, conflict, tension, cyprus, ireland, end, proxi
Political Authority	administr, trump, they, presid, inde, parti, progress, leadership, republican, easi
Political Corruption	corrupt, money, lobbyist, lobbi, special, kleptocrat, ban, pac, disclosur, anticorrupt
Protectionism	trade, currenc, export, competit, ship, unfair, corpor, compet, domest, open
Social Groups	class, veteran, middl, politician, care, young, youngster, lifelin, honor, cemeteri
Sustainability	climat, environment, environ, conserv, pollut, clean, green, ocean, agre, both
Technology and Infrastructure	research, technolog, infrastructur, transport, innov, broadband, highway, grid, scienc, train
Traditional Morality	abort, marriag, famili, faith, religi, parent, faithbas, life, first, marri
Welfare State	health, hous, medicar, healthcar, medicaid, poor, servic, va, coverag, charit

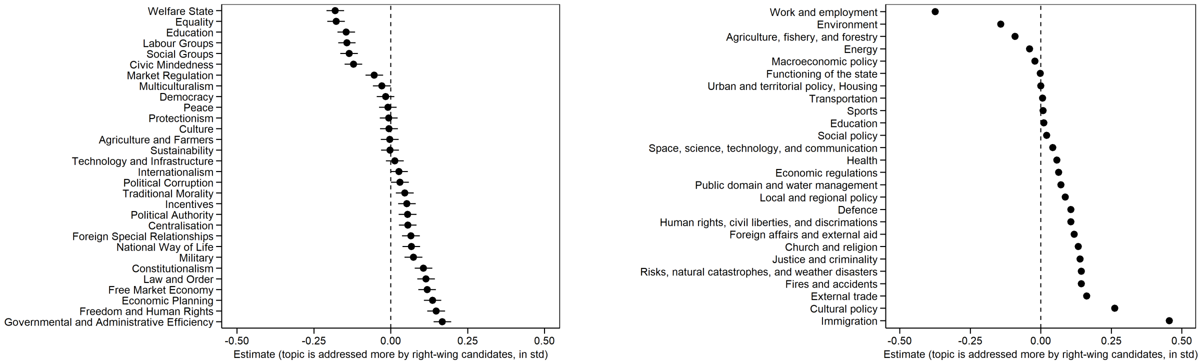
(b) France

Topics	Most predictive stemmed words
Translation	*Translation*
Affaires internationales et aide extérieure	leurop, europeen, diplomat, europ, trait, chin, sud, procheorient, pacif, mediterraneen
Foreign affairs and external aid	*the Europe, European, diplomacy, Europe, treaty, china, south, Middle East, pacific, Mediteranean*
Agriculture pêche et sylviculture	agricol, agricultur, agriculteur, pech, lagricultur, alimentair, ogm, animal, paysan, dagriculteur
Agriculture, fishery, and forestry	*agricultural, agriculture, farmer, fishing, the agriculture, alimentary, GMO, animal, peasant, for farmers*
Autres	laven, vert, programm, sacr, rassembl, tach, march, lecolog, gauch, letat
Other	*before, green, program, sacred, gather, task, walk, environment, left, state*
Commerce extérieur	mondialis, competitivit, commercial, libreechang, lexport, echang, reciprocit, exterieur, douan, export
External trade	*globalisation, competitivity, comercial, free trade, the export, trade, reciprocity, external, customs, export*
Défense	defens, arme, militair, darm, guerr, lotan, paix, desarm, larme, combatt
Defence	*defence, weapon, military, army, war, NATO, peace, disarmament, the military, combat*
Domaine public et gestion de l eau	leau, sacrifie, paysag, eglis, foret, autrefois, priv, remunicipalis, tol, hangar
Public domain and water management	*water, sacrified, landscape, church, forest, past, private, remunicipalisation, barn, toll*
Droits de l homme libertés publiques et discriminations	discrimin, femm, laicit, handicap, respect, libert, luniversalit, legalit, lib, religi
Human rights, civil liberties, and discrimations	*discrimination, women, secularism, handicap, respect, freedom, universality, equality, liberty, religion*
Education	scolair, leduc, lecol, enseign, lenseign, elev, ecol, educ, universit, etudi
Education	*academic, the education, school, teaching, student, school, education, university, student*
Eglises et religion	faim, mediat, vient, exact, celuic, normal, reintroduir, contemporain, voit, est
Church and religion	*hunger, mediation, come, exact, clergy, normal, reintroduction, contemporary, see, is*
Energie	nucleair, energet, denerg, lenerg, energ, gazol, electr, renouvel, fessenheim, lelectricit
Energy	*nuclear, energetic, energy, gas, electricity, renewable, fessenheim (a nuclear power plant), the electricity*
Environnement	ecolog, lenviron, environnemental, naturel, dechet, biodiversit, natur, eau, pollut, environ
Environment	*ecology, environment, environmental, natural, waste, biodiversity, nature, water, pollution, surroundings*
Espace Science Technologie et communication	numer, recherch, technolog, internet, telecommun, logiciel, ntic, scienc, chercheur, dadvs
Space, science, technology, and communication	*numeric, research, technology, internet, telecommunication, software, ICT, science, researchers, DADVS law*
Fonctionnement de l Etat	decentralis, fonctionnair, privatis, administr, local, referendum, corrupt, nationalis, public, ministr
Functioning of the state	*decentralisation, public servant, privatisation, administration, local, referendum, corruption, nationalisation, public, ministry*
Immigration	limmigr, immigr, nationalit, clandestin, migratoir, naturalis, dasil, appliquonsl, dimmigr, migrat
Immigration	*immigration, nationality, clandestine, migratory, naturalisation, asile, apply, migration*
Incendies et accidents	faim, pompi, sapeur, paysag, eglis, sacrifie, leau, priv, competit, scientif
Fires and accidents	*hunger, firefighters, sapper (firefighters), landscape, church, sacrifice, water, private, competition, scientific*
Justice et Criminalité	delinqu, polic, justic, violenc, victim, magistrat, jug, prison, judiciair, securit
Justice and criminality	*delinquency, police, justice, violence, victim, magistrate, judge, prison, judiciary, security*
Politique culturelle	culturel, cultur, audiovisuel, artist, francophon, laudiovisuel, langu, medi, press, televis
Cultural policy	*cultural, culture, audiovisual, artist, francophone, the audiovisual, language, medical, median, tv*
Politique locale et régionale	region, loutrem, regional, doutrem, domtom, caledon, elementcl, memoir, aerien, concerne
Local and regional policy	*region, overseas, regional, from overseas, french islands, New Caledonia, element, memory, aerial, concern*
Politique Macroéconomique	fiscal, leuro, fiscalit, det, croissanc, impot, econom, limpot, industriel, budgetair
Macroeconomic policy	*fiscal, euro, fiscality, debt, growth, tax, economy, taxes, industrial, budgetary*
Politique sociale	associ, familial, jeuness, famill, crech, rmi, vieilless, vacanc, proposis, inegalit
Social policy	*association, familial, youth, family, day care, minimum wage, ageing, holidays, proposition, inequality*
Politiques urbaines et territoriales Logement	log, quarti, propriet, ruralit, vill, locat, loyer, rural, egalit, dheberg
Urban and territorial policy, Housing	*housing, neighborhood, landlord, rurality, city, tenant, rent, rural, equality, housing*
Régulations économiques	pme, pmepm, lartisanat, specul, independ, tourist, cred, lentrepreneuriat, bancair, banqu
Economic regulations	*SME, MSME, craft, speculation, independance, touristic, credibility, entrepreneurship, banking, bank*
Risques et catastrophes naturels et météorologiques	daccident, conven, popul, nuisanc, faim, paysag, eglis, sacrifie, concerne, priv
Risks, natural catastrophes, and weather disasters	*accident, convention, popular, disturbance, hunger, lanscape, church, sacrifice, concern, private*
Santé	sant, medecin, medical, soin, medic, lhopital, prevent, sanitair, malad, lassurancemalad
Health	*health, doctor, medical, care, medical, hospital, preventive, sanitary, sickness, health insurance*
Sports	sportif, sportiv, sport, athlet, competit, dopag, fiert, reconnaiss, pratiqu, haut
Sports	*sport, sportsman, sport, athlete, competition, doping, pride, recognition, practice, high*
Transport	routi, transport, ferroviair, maritim, rail, infrastructur, navir, pavillon, flott, voitur
Transportation	*road, transport, rail, maritime, railway, infrastructure, ship, flag, fleet, car*
Travail et emploi	travail, retrait, travailleur, syndical, salair, salar, heur, professionnel, travaill, syndicat
Work and employment	*work, retirement, worker, union, wage, salary, hour, professional, worker, union*

Notes: We list the most predictive stemmed words associated with each topic in the U.S. party manifestos coded by the Manifesto Project (Panel a) and in the French party manifestos coded by the Agenda Project (Panel b).

Validation against ideological scores In order to validate our topic measure, we regress the intensity of each topic against candidates' ideological scores, while controlling for year fixed effects. In the U.S., Democratic candidates are more likely to address issues such as equality, the welfare state, and labour groups, while Republican candidates are more likely to cover subjects such as the free market economy, freedom and human rights, and administrative efficiency (Figure B.3a). In France, left-wing candidates are more likely to address environmental issues and employment, and right-wing candidates are more likely to address immigration and cultural policy (Figure B.3b).

Figure B.3: Topics related to candidates' ideology

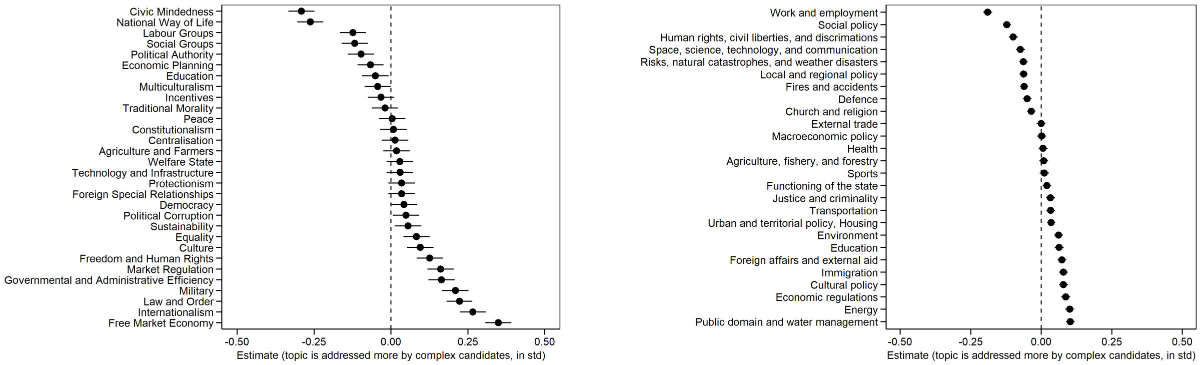

(a) U.S.

(b) France

Notes: For each topic, we show the standardized point estimate and 95% confidence interval from a regression of the prevalence of that topic in a candidate website in the U.S. (Panel a) and a candidate manifesto in France (Panel b) on the website or manifesto's ideological score, controlling for year fixed effects. We use one observation per candidate and election round and the sample includes all available general election website captures (N=5,792) and manifestos (N=56,915) for which we could compute textual metrics.

Validation against complexity. Similarly, we regress the intensity of each topic against candidates' complexity, including year fixed effects. Both in the U.S and in France, candidates with a more complex platform are more likely to address topics such as the economic regulation and the environment, while less complex candidates are more likely to discuss the functioning of the state and labour groups (Figures B.4a and B.4b) .

Figure B.4: Topics related to candidates' complexity


(a) U.S.

(b) France

Notes: For each topic, we show the standardized point estimate and 95% confidence interval from a regression of the prevalence of that topic in a candidate website in the U.S. (Panel a) and a candidate manifesto in France (France) on the website or manifesto's complexity score, controlling for year fixed effects. Other notes as in Figure B.3.

B.5 Vector representations

Many techniques enable to transform texts into vectors (also called text embeddings) in a multidimensional space. We use some of the most widespread, from basic to state-of-the-art:

- TF-IDF (term frequency-inverse document frequency): texts are first converted into a term frequency matrix. Each row represents a text i, each column represents a word j available in the entire corpus of texts, and each cell indicates the frequency $f_{i,j}$ of word j in text i. Words are then inversely weighed based on the number of texts in which they appear. Indeed, frequent words are less likely to carry meaning and discriminate texts. For instance, words such as "a," "the," etc. are likely to appear in many texts, and do not differentiate them. On the contrary, words such as "medicare" appear in fewer texts and should receive a larger weight when assessing text similarity. For text $i \in T$, the TF-IDF representation across words j is $\left(f_{i,j}/\sum_{t\in T} f'_{i,t}\right)_j$.

- LSI (latent semantic indexing): even excluding very frequent and very rare words (which are often typos), TF-IDF vectors reach high dimensions, typically 5,000-10,000 in our corpus. In high dimension, vectors tend to become sparse and appear dissimilar from any other, a phenomenon known as "Curse of Dimensionality." To mitigate this problem, latent semantic indexing performs a singular value decomposition of TF-IDF matrices and only keeps the highest variance bearing dimensions (typically around 100).

- W2V (word2vec): word2vec relies on a neural network trained to predict the next word given the beginning of a sentence. During the training process, word2vec implicitly creates *word* embeddings, i.e., word vectorial representations. Since word2vec creates embeddings using words' contexts, it is arguably better able to group words with similar meanings than TF-IDF. Once word2vec is trained, we can calculate text representations by taking the average of word embeddings. Given the large size of our U.S. and French corpora, we are able to train word2vec models ourselves. To improve the quality of our training, we use all available manifestos and websites. We also rely on pre-trained models provided by Yamada et al. (2020) and Fauconnier (2015) for English and French content respectively.

- BERT (Bidirectional Encoder Representations from Transformers): BERT models are state-of-the-art for many NLP applications. Similarly as word2vec, BERT also works around word embeddings. The main difference is that BERT generates *context dependent* word embeddings. For example, in the sentence "I left my coat on the left side of the room," the word "left" would have a single embedding in word2vec – a combination of all meanings of "left" – whereas it would have two different word embeddings in BERT. BERT models require vast amounts of training data to perform well. Hence, we rely on pre-trained models exclusively: *bert-base-*

uncased by Devlin et al. (2018) for English content and *camembert-base* by Martin et al. (2020) for French content.

To test the quality of our text embeddings, we conduct a series of prediction tasks and assess the accuracy of each embedding model. Specifically, we attempt to predict candidates' characteristics (e.g., their party, region, gender, and incumbency status) based on the text embedding of their website or manifesto, fed into a logistic regression. In the U.S., the LSI and TF-IDF models perform best on average, followed by BERT and W2V. In France, BERT and W2V perform best, followed by LSI and TF-IDF. That said, different models perform differently for different tasks: the standard deviation of models' ranking across tasks is about 2. For that reason, and not knowing a priori which model is better suited to identify text similarity, we use an index of all models.

B.6 Text similarity

Given two vectorial representations of texts x and y, one can then calculate the cosine similarity:

$$similarity(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|} = \frac{\sum\limits_i x_i y_i}{\sqrt{\sum\limits_i x_i^2 \sum\limits_i y_i^2}}.$$

In a two-dimensional space, the cosine similarity is the analog of the cosine of the angle between vectors x and y. By definition, the cosine similarity is included between -1 and 1. When x and y point toward the exact same direction (resp. the opposite direction), the cosine similarity equals 1 (resp. -1). The cosine similarity does not depend on vectors' norms, i.e. text lengths. This makes it possible to compare texts of different lengths using the cosine similarity, whereas under the Euclidian distance, small texts and large texts would be deemed very different regardless of their content. In practice, all our calculations of similarity fall between 0 and 1.

C Additional Results on the Convergence between Rounds

Table C.1: Candidates' ideology related to voters' characteristics

	Candidates' ideology	
	U.S.	France
	(1)	(2)
Log median income	.389***	−.207**
	(.101)	(.100)
Share of foreign born	−.688**	−.530**
	(.278)	(.260)
Log population density	−.011	−.051***
	(.017)	(.014)
Share of high school diploma	−1.005*	−2.011***
	(.580)	(.679)
Unemployment	−1.475	.283
	(1.256)	(.473)
Population median age	.003	.009**
	(.005)	(.004)
Previous presidential right-wing vote	.712***	1.064***
	(.228)	(.164)
Year FE	X	X
Observations	1,870	1,791
R^2	.023	.125

Notes: Standard errors, shown in parentheses, are clustered by district × year and we indicate significance at 1, 5, and 10% with ***, **, and *, respectively. The outcome is the ideological score of a candidate website in the U.S. (column 1) or a candidate manifesto in France (column 2). We use one observation per candidate, and the sample includes candidates competing in a general election in 2010 and later (column 1) and candidates competing in a second round in 2007 and later (column 2). Voter characteristics are measured at the constituency level and we control for year fixed effects.

Table C.2: Candidates' complexity related to voters' characteristics

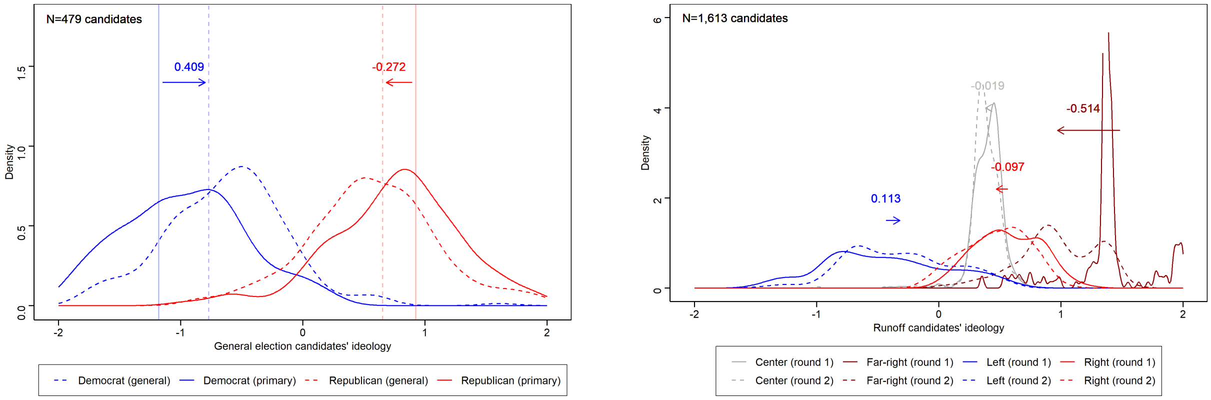
	Candidates' complexity	
	U.S.	France
	(1)	(2)
Log median income	.002	−.154
	(.074)	(.110)
Share of foreign born	−.300	−.590*
	(.204)	(.334)
Log population density	.014	.009
	(.013)	(.015)
Share of high school diploma	.686*	1.083
	(.357)	(.828)
Unemployment	.345	−.134
	(1.151)	(.547)
Population median age	−.007*	−.004
	(.004)	(.004)
Previous presidential right-wing vote	−.330**	−.027
	(.163)	(.203)
Year FE	X	X
Observations	1,870	1,791
R^2	.012	.054

Notes: The outcome is the complexity score of a candidate website in the U.S. (column 1) or a candidate manifesto in France (column 2). Other notes as in Table C.1.

Figure C.1: Ideology moderation (second round with exactly two candidates)

(a) U.S.

(b) France

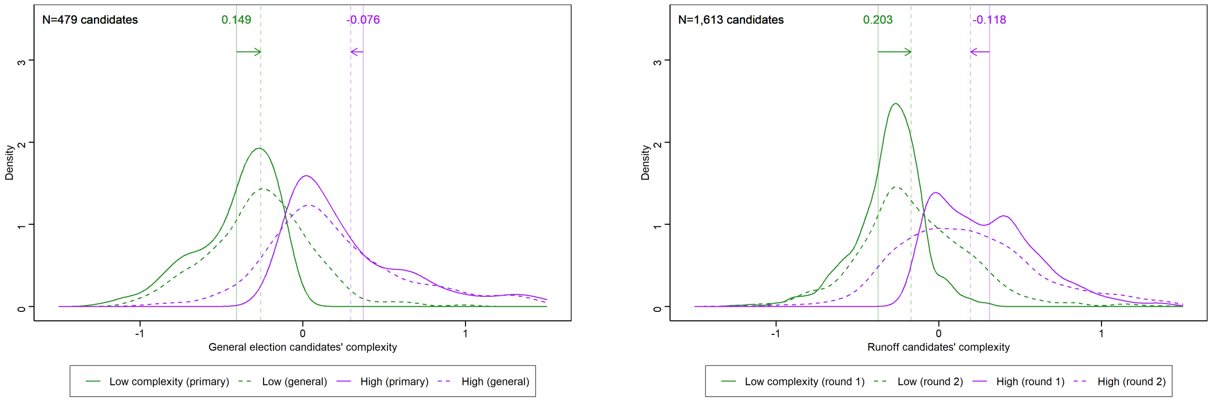


Notes: The sample is restricted to races where exactly two candidates, the leader and the qualified opponent, are present in the second round. Specifically, we exclude general elections where third-party candidates are present and where a primary election winner drops out before the general election, in the U.S.; and runoffs where more than two candidates qualify for the second round, as well as runoffs where only two candidates qualify but one of them drops out of the race, in France. Other notes as in Figure 1.

Figure C.2: Complexity moderation (second round with exactly two candidates)

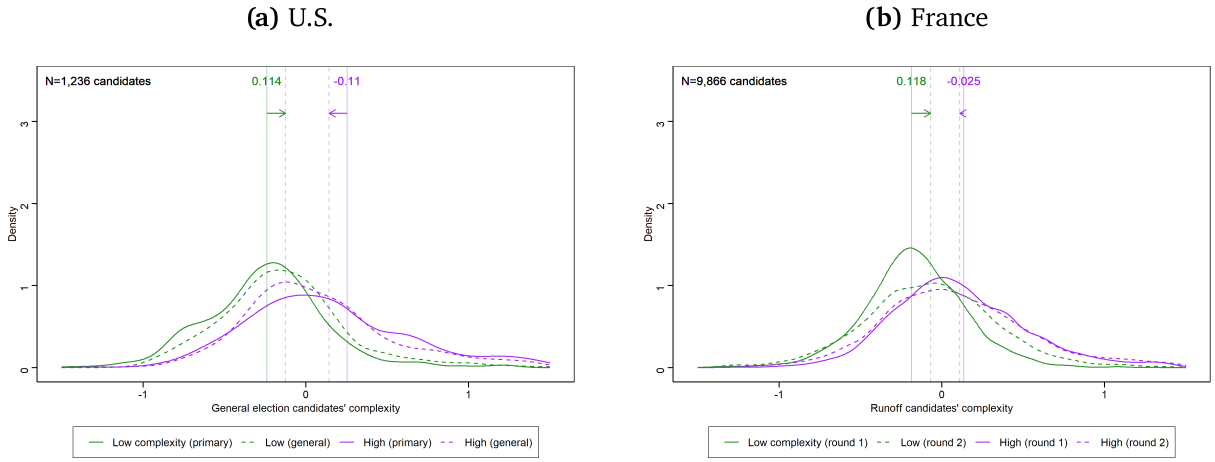
(a) U.S.

(b) France



Notes: Same notes as in Figures 2 and C.1.

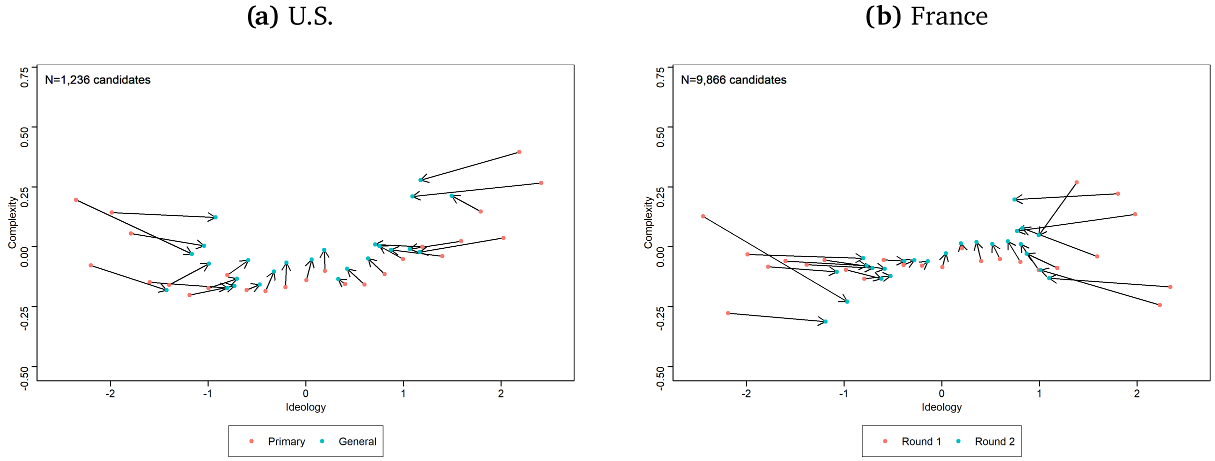
Figure C.3: Complexity moderation (split based on predicted complexity)


(a) U.S.

(b) France

Notes: We plot the kernel density of candidates' complexity score, pooling all election years together and splitting the sample between candidates whose predicted complexity score in the first round is below the median score in a given election year, and those whose predicted complexity score is above the median. We predict complexity based on district fixed effects and candidate-specific variables: the candidate's party or political orientation, whether they are the incumbent, whether their party or political orientation won the previous election, their party or political orientation's vote share in the previous election, and the length of their website or manifesto. Other notes as in Figure 1.

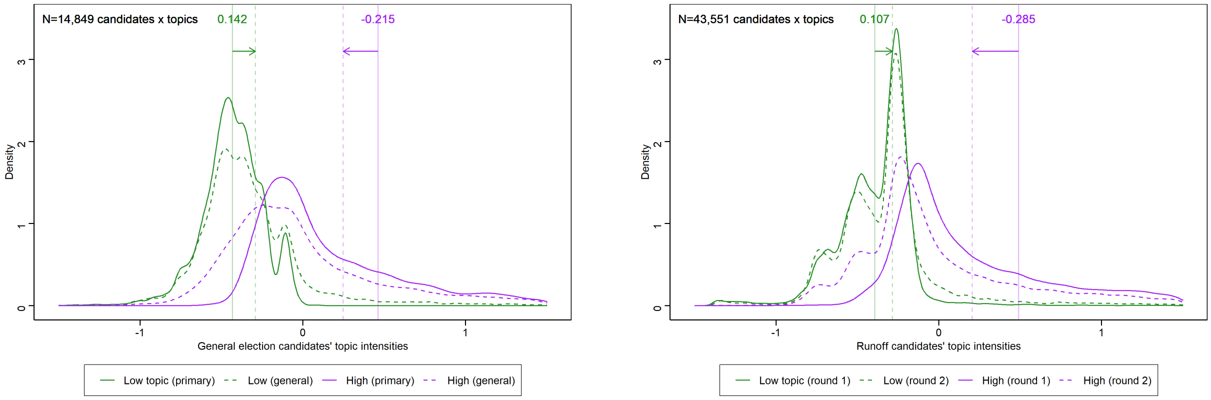
Figure C.4: Ideology-complexity joint moderation


(a) U.S.

(b) France

Notes: We plot the mean complexity score against the mean ideology score within each bin of the ideology score in the primary election or the first round, as well as each bin's corresponding mean complexity and mean ideology in the general election or the second round. Other notes as in Figure 1.

Figure C.5: Topics moderation (second round with exactly two candidates)

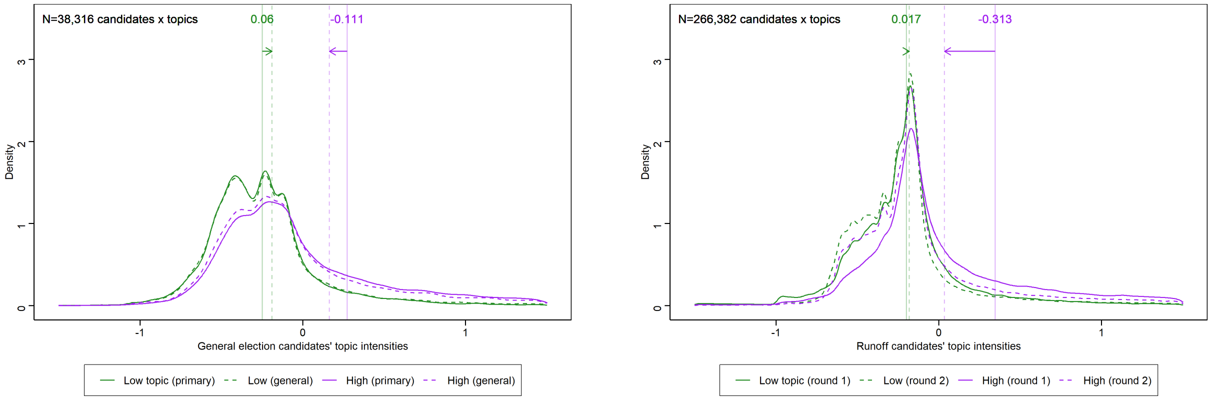


(a) U.S.

(b) France

Notes: Same notes as in Figures 3 and C.1.

Figure C.6: Topics moderation (split based on predicted topic propensities)



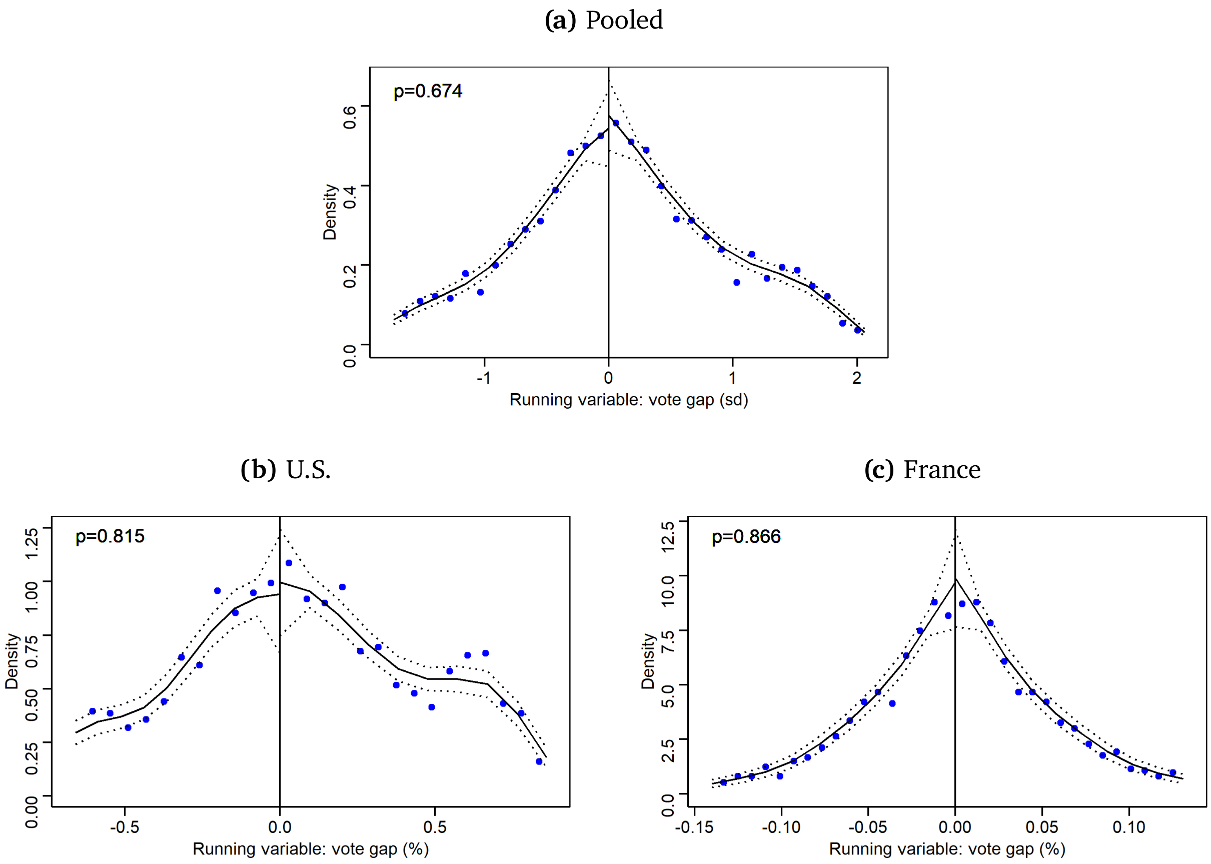
(a) U.S.

(b) France

Notes: We plot the kernel density of candidates' topic prevalence, pooling all election years and topics together. For each topic, we split the sample between candidates whose predicted topic prevalence in the first round is below the median predicted topic prevalence in a given election year, and those whose topic prevalence is above the median. We predict topic propensities based on district fixed effects and candidate-specific variables: the candidate's party or political orientation, whether they are the incumbent, whether their party or political orientation won the previous election, their party or political orientation's vote share in the previous election, and the length of their website or manifesto. Other notes as in Figure 3.

D Additional Results on the Adjustment to Opponent

Figure D.1: McCrary balance tests



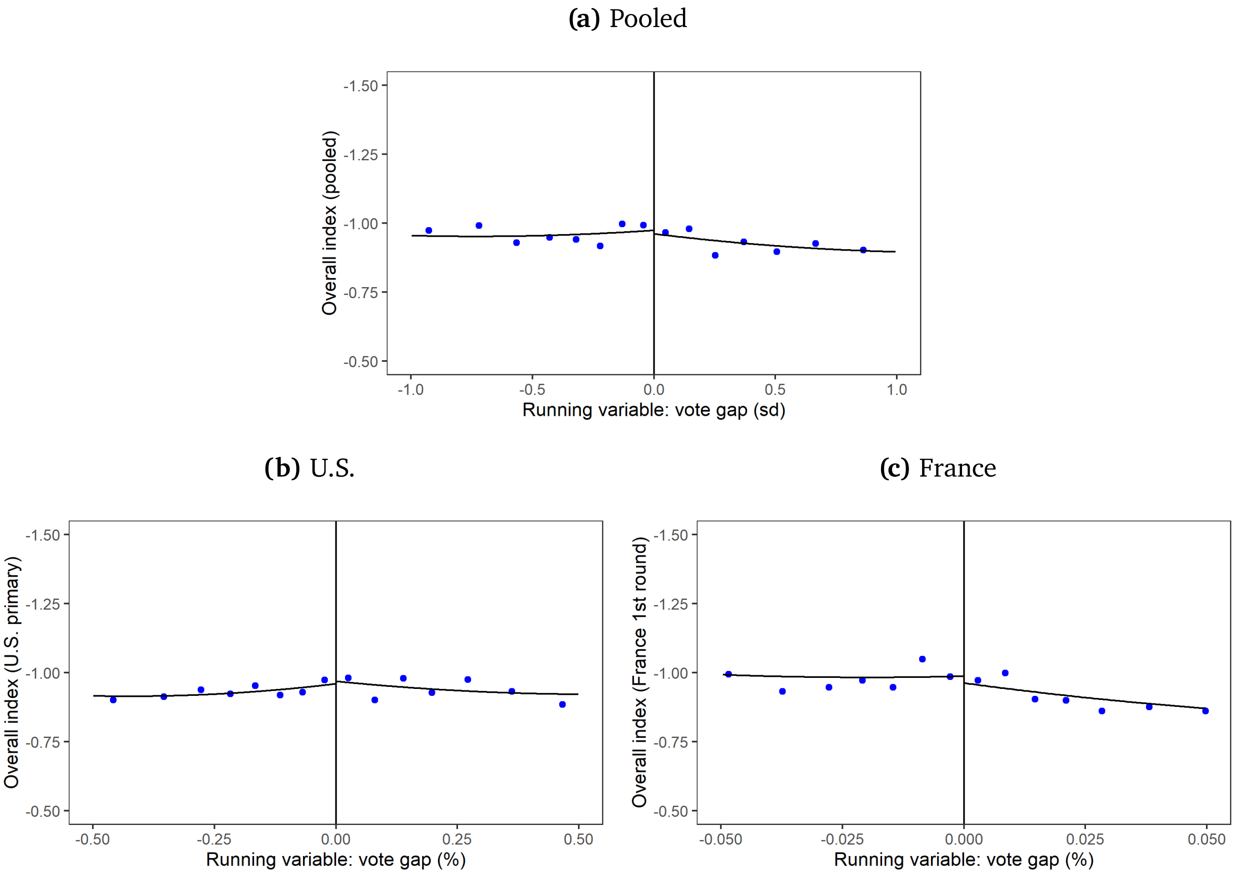
(a) Pooled

(b) U.S.

(c) France

Notes: This figure tests if there is a jump at the threshold in the density of the running variable: the vote share difference between the top two candidates in the primary election, in the U.S. (Figure D.1b); and the vote share difference between the second- and third-ranked candidates in the first election round, in France (Figure D.1c). The solid curve is a quadratic fit and the confidence intervals are represented by dashed curves. We use the R implementation of the rddensity package (Cattaneo, Jansson and Ma, 2018) to create the charts and to compute p-values.

Figure D.2: Overall similarity in the first round



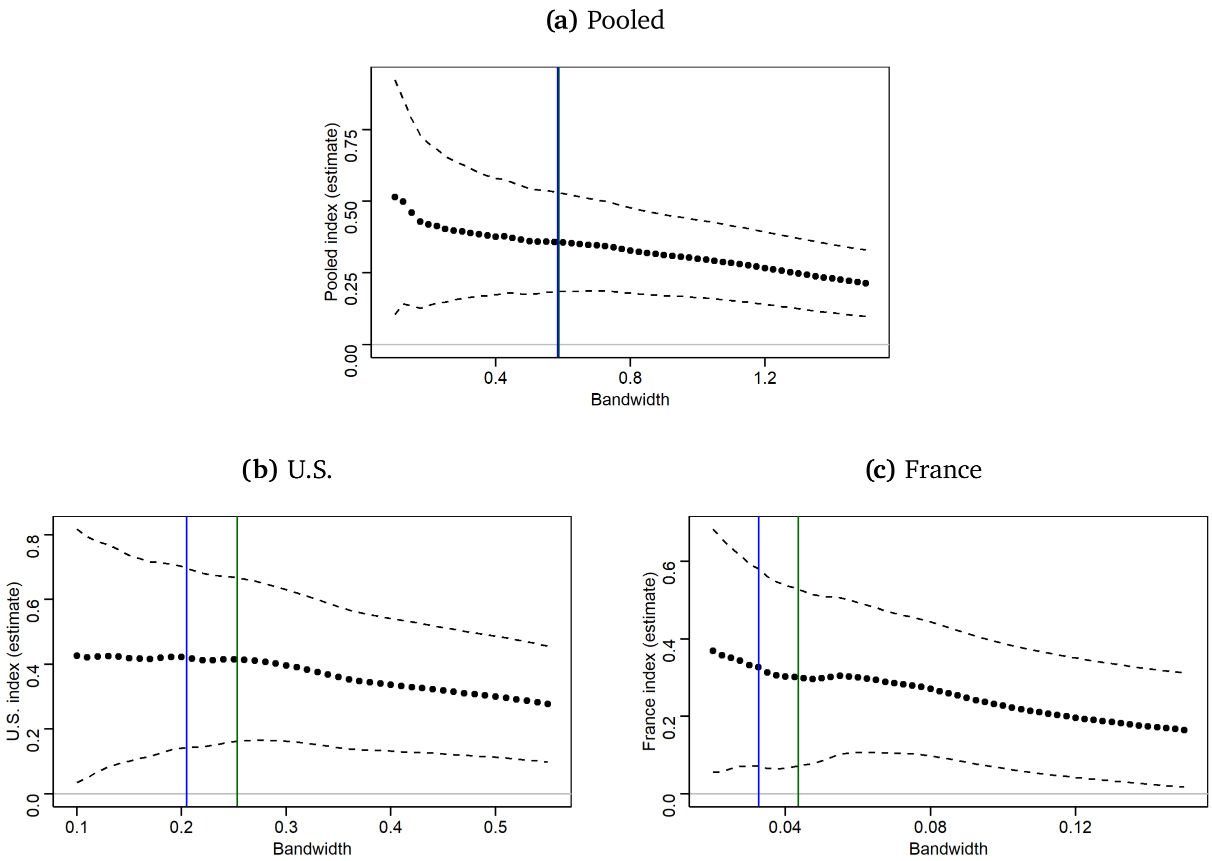
(a) Pooled

(b) U.S.

(c) France

Notes: The outcome is the overall similarity between the candidate's website or manifesto and that of the leader in the primary election (U.S.) or the first round (France). It is defined as the average of the cosine similarity between vectorized texts (standardized and averaged across all vector representations), negative the distance between ideology scores, negative the distance between complexity measures (standardized and averaged across the three measures of complexity), and negative the Euclidean distance between topic distributions. It is constructed separately and divided by its standard deviation within each country. Other notes as in Figure 4.

Figure D.3: Bandwidth robustness



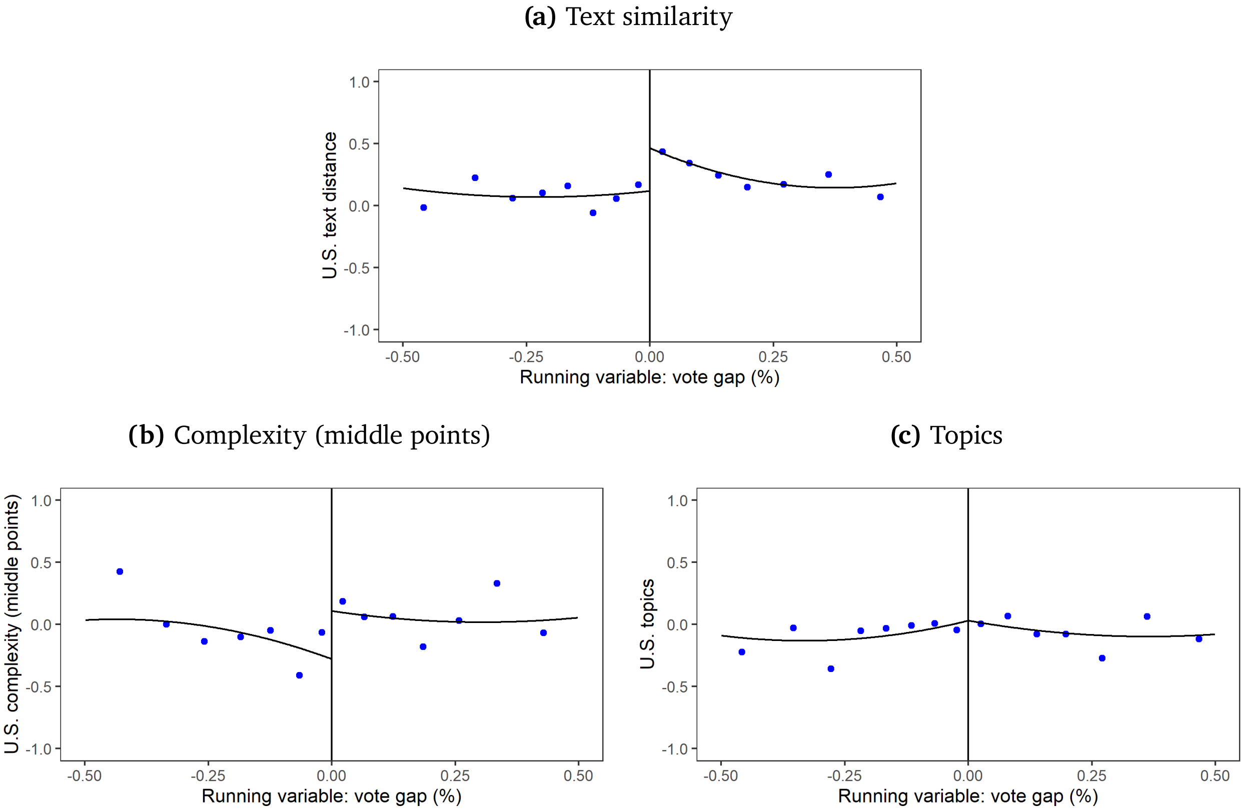
(a) Pooled

(b) U.S.

(c) France

Notes: This figure tests the robustness of our main estimate to several bandwidths. The optimal bandwidth chosen by the MSERD procedure from Calonico et al. (2019) is indicated with a green line and the optimal bandwidth chosen by the IK procedure from Imbens and Kalyanaraman (2012) is indicated with a blue line. The outcome is the overall similarity to the winner, defined as the average of the (standardized) vectorized text similarity, negative the distance in ideology, negative the distance in complexity, and negative the distance in topic distribution. Other notes as in Figure 4.

Figure D.4: Convergence on different dimensions in the U.S.



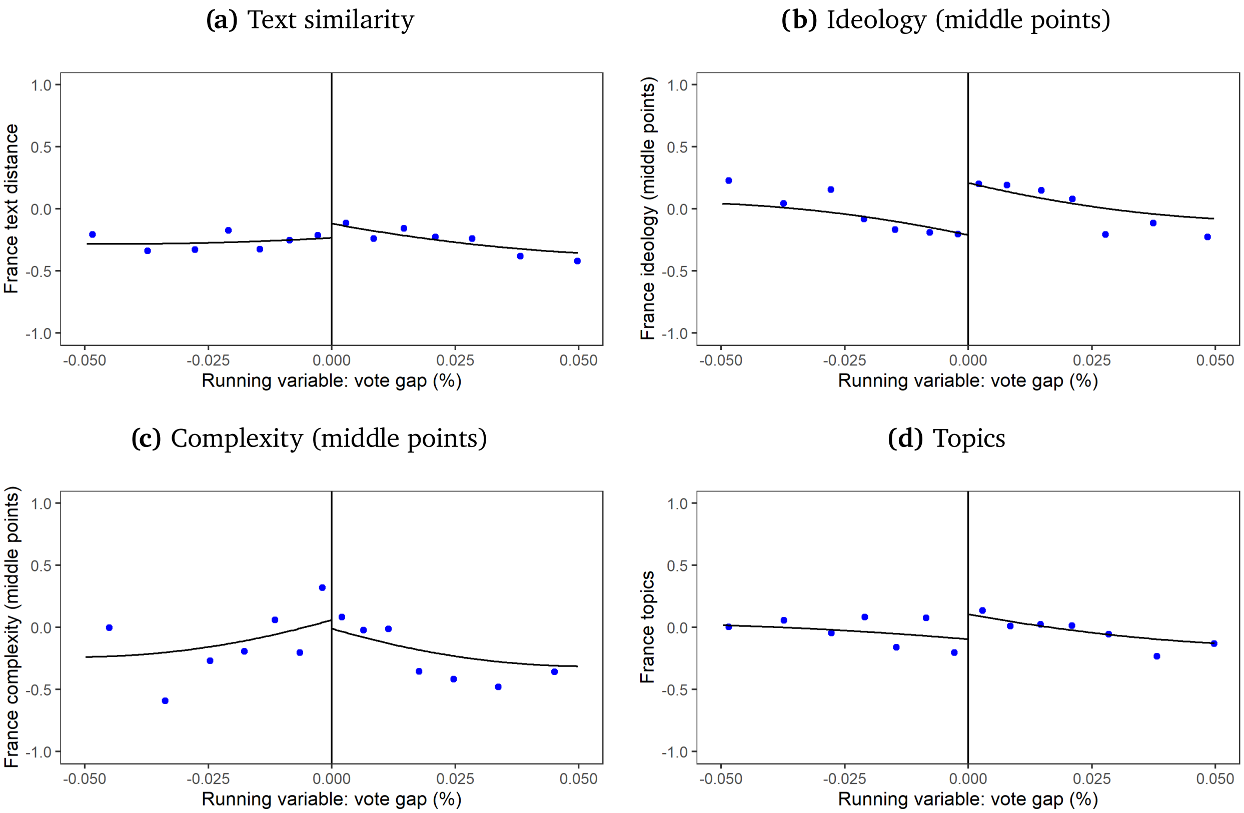
(a) Text similarity

(b) Complexity (middle points)

(c) Topics

Notes: The outcome is the change in similarity to the opponent or runner-up between election rounds, in terms of vectorized text similarity (Figure D.4a), complexity score (Figure D.4b), and topic distribution (Figure D.4c). In Figure D.4b, the sample is restricted to races in which the leader is initially in the middle of the two possible opponents on the complexity scale.

Figure D.5: Convergence on different dimensions in France



(a) Text similarity

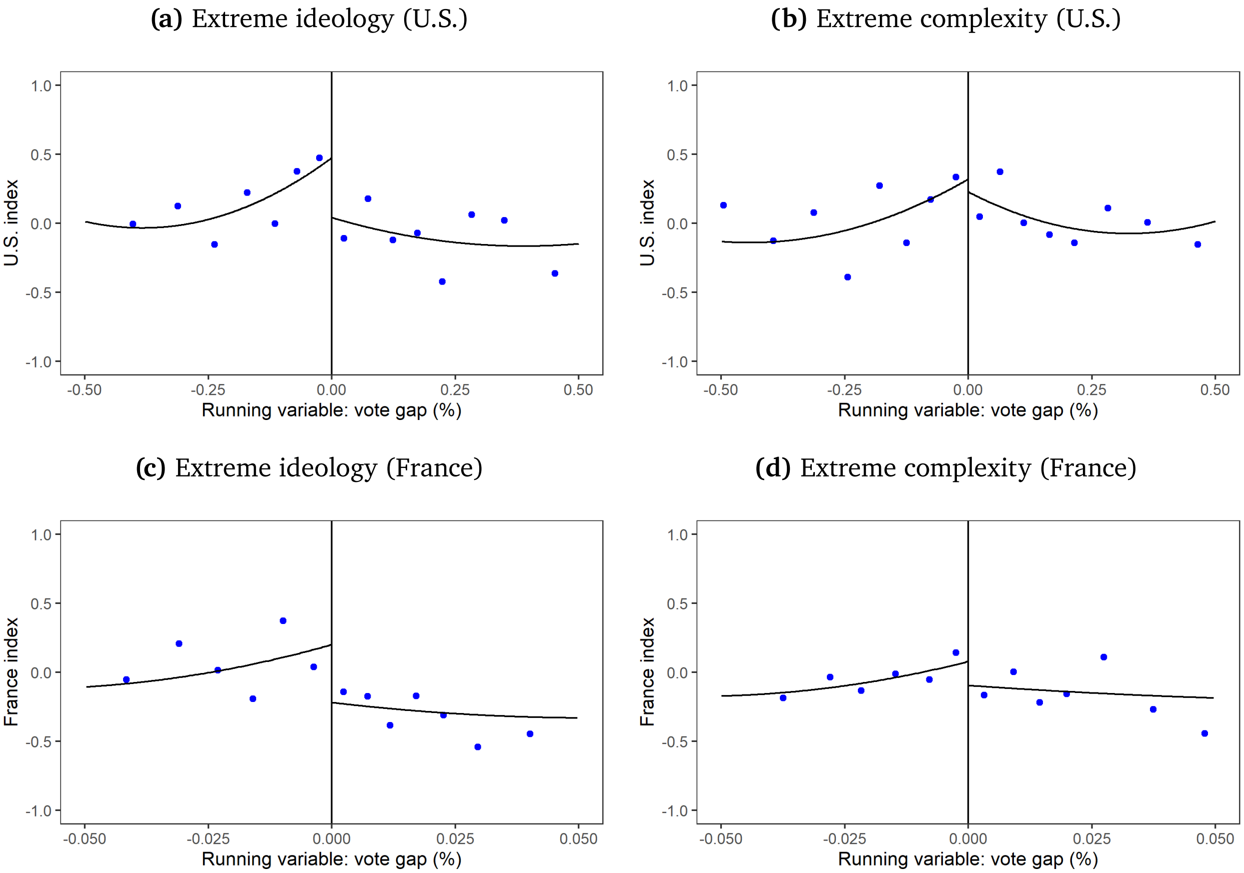
(b) Ideology (middle points)

(c) Complexity (middle points)

(d) Topics

Notes: In Figure D.5b, the sample is restricted to races in which the leader is initially in the middle of the two possible opponents on the ideology scale. Other notes as in Figure D.4.

Figure D.6: Convergence to the more extreme opponent



(a) Extreme ideology (U.S.)

(b) Extreme complexity (U.S.)

(c) Extreme ideology (France)

(d) Extreme complexity (France)

Notes: The outcome is the change in overall similarity between the leader and the opponent between election rounds, defined as the average of the standardized changes in vectorized text similarity as well as similarity in ideology, complexity, and topic distribution. It is constructed separately and divided by its standard deviation within each country. There is one observation per race and the running variable is defined as the difference in vote shares between the more extreme candidate (either in terms of ideology or complexity) and the other potential opponent. It is positive in races where the qualified opponent is more extreme and negative in races where the qualified opponent is more moderate. Other notes as in Figure 4.

Table D.1: U.S. regression discontinuity sampling frame

Year	Races in RD Sample	Mean # of Candidates	Mean Qualifying Margin
2002	68	3.3	18 %
2004	103	2.7	26 %
2006	127	3.1	25 %
2008	134	2.8	24 %
2010	217	3	23 %
2012	208	2.8	30 %
2014	186	2.7	28 %
2016	182	2.9	27 %

Notes: For each election at the U.S. House of Representatives, we indicate the number of primary races included in the regression discontinuity design, the average number of candidates in these races, and the average qualification margin, defined as the difference in vote share between the primary winner and the closest contender.

Table D.2: France regression discontinuity sampling frame

(a) Parliamentary elections

Year	Races in RD Sample	Mean # of Candidates	Mean Qualifying Margin
1978	1	17	1.4 %
1981	1	7	2 %
1988	4	6.2	3.3 %
1993	162	9.8	3.1 %
1997	28	12.6	2.8 %
2002	3	17.3	2.3 %
2007	2	15	5.5 %
2012	1	11	12.9 %
2017	221	13.8	2.8 %
2022	310	10.9	3.5 %

(b) Local elections

Year	Races in RD Sample	Mean # of Candidates	Mean Qualifying Margin
1985	2	7	0.9 %
1988	12	6	4.3 %
2001	6	7.3	3.4 %
2008	1	6	2.4 %
2011	36	6.8	6.2 %
2015	8	5.8	5.3 %
2021	9	3.4	8.5 %

Notes: For each French parliamentary and local election, we indicate the number of first rounds included in the regression discontinuity design, the average number of candidates in these races, and the average qualification margin, defined as the difference in vote shares between the second- and third-ranked candidates in the first round.

Table D.3: Balance tests (part 1)

(a) U.S.

Outcome	Available (1)	Democrat (2)	Male (3)	Website length (4)
Treatment	0.027	-0.020	0.013	-129.9
	(0.042)	(0.068)	(0.063)	(157.2)
Robust p-value	0.492	0.717	0.658	0.348
Observations left	2708	821	821	821
Observations right	2708	1031	1031	1031
Effective obs. left	1109	416	325	429
Effective obs. right	1109	429	347	443
Polyn. order	1	1	1	1
Bandwidth	0.242	0.246	0.191	0.255
Mean, left of threshold	0.303	0.443	0.177	416.359

(b) France

Outcome	Available (1)	Left-wing (2)	Male (3)	Manifesto length (4)
Treatment	0.005	-0.000	-0.015	-46.9
	(0.013)	(0.064)	(0.070)	(74.7)
Robust p-value	0.664	0.929	0.879	0.494
Observations left	6923	688	677	688
Observations right	6923	721	706	721
Effective obs. left	4592	488	351	382
Effective obs. right	4592	500	359	393
Polyn. order	1	1	1	1
Bandwidth	0.065	0.055	0.033	0.038
Mean, left of threshold	0.100	0.420	0.363	738.599

Notes: In column 1, the outcome is a dummy equal to 1 if the website or manifesto of the leader is available in both election rounds and if the primary election website (Panel a) or first-round manifesto (Panel b) of the candidate (opponent or runner-up) is available. In column 2, the outcome is a dummy equal to 1 if the candidate is a Democrat (Panel a) or on the left or the far-left (Panel b). In column 3, the outcome is a dummy equal to 1 if the candidate is a man. In column 4, the outcome is the number of words in the candidate's website (Panel a) or manifesto (Panel b). In column 1, the sample includes all candidates running in a competitive primary (Panel a) or a first round where the second-ranked candidate does not pass the runoff qualification threshold (Panel b). In all other columns, the sample is the RDD sample described in Section 5.1. Other notes as in Table 1.

Table D.4: Balance tests (part 2)

(a) U.S.

Outcome	Ran before (1)	Incumbent (2)	Party is incumbent (3)
Treatment	-0.005	0.005	-0.037
	(0.071)	(0.036)	(0.066)
Robust p-value	0.880	0.696	0.557
Observations left	607	607	607
Observations right	773	773	773
Effective obs. left	267	321	361
Effective obs. right	293	346	395
Polyn. order	1	1	1
Bandwidth	0.213	0.268	0.321
Mean, left of threshold	0.137	0.015	0.420

(b) France

Outcome	Ran before (1)	Incumbent (2)	Orientation is incumbent (3)
Treatment	0.046	0.017	0.198***
	(0.061)	(0.042)	(0.076)
Robust p-value	0.358	0.664	0.007
Observations left	662	662	644
Observations right	695	695	680
Effective obs. left	368	431	289
Effective obs. right	378	444	297
Polyn. order	1	1	1
Bandwidth	0.038	0.048	0.027
Mean, left of threshold	0.204	0.066	0.250

Notes: In column 1, the outcome is a dummy equal to 1 if the candidate (opponent or runner-up) ran in the previous election in the same district. In column 2, the outcome is a dummy equal to 1 if the candidate won the previous election in the same district. In column 3, the outcome is a dummy equal to 1 if the previous election was won by the candidate's party (Panel a) or orientation (Panel b). The sample excludes candidates running in districts that were redistricted since the previous election. Other notes as in Table 1.

Table D.5: Balance tests (part 3)

(a) U.S.

Outcome	Index (1)	Similarity (2)	Ideology (3)	Complexity (4)	Topics (5)
Treatment	-0.001	-0.024	0.082	0.044	-0.092
	(0.047)	(0.052)	(0.079)	(0.085)	(0.056)
Robust p-value	0.975	0.691	0.327	0.555	0.157
Observations left	821	821	821	821	821
Observations right	1031	1031	1031	1031	1031
Effective obs. left	495	483	460	399	404
Effective obs. right	516	506	478	414	417
Polyn. order	1	1	1	1	1
Bandwidth	0.315	0.304	0.283	0.230	0.235
Mean, left of threshold	-0.929	-0.893	-0.838	-0.800	-0.927

(b) France

Outcome	Index (1)	Similarity (2)	Ideology (3)	Complexity (4)	Topics (5)
Treatment	0.031	0.035	0.067	-0.008	0.014
	(0.057)	(0.068)	(0.067)	(0.084)	(0.068)
Robust p-value	0.615	0.625	0.327	0.996	0.891
Observations left	688	688	688	688	688
Observations right	721	721	721	721	721
Effective obs. left	378	371	448	379	381
Effective obs. right	384	379	461	387	390
Polyn. order	1	1	1	1	1
Bandwidth	0.036	0.035	0.048	0.037	0.037
Mean, left of threshold	-1.001	-1.044	-0.810	-0.763	-1.001

Notes: The outcome is the similarity between the leader's website (Panel a) or manifesto (Panel b) and that of the candidate (opponent or runner-up) in the primary election (Panel a) or the first round (Panel b), in terms of vectorized text similarity (column 2), ideological score (column 3), complexity score (column 4), and topic distribution (column 5). In column 1, the outcome is the average of these four (standardized) measures of similarity. Each outcome is divided by its standard deviation. Other notes as in Table 1.

Table D.6: Convergence on different dimensions, controlling for unbalanced covariate

(a) U.S.

Outcome	Index	Text similarity	Ideology	Complexity		Topics
	Full sample	Full sample	Full sample	Full sample	Middle points	Full sample
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.409**	0.405**	0.155	0.350**	0.453*	0.075
	(0.154)	(0.154)	(0.143)	(0.159)	(0.251)	(0.132)
Robust p-value	0.012	0.017	0.258	0.042	0.093	0.625
Observations left	821	821	821	821	207	821
Observations right	1031	1031	1031	1031	207	1031
Effective obs. left	421	408	404	444	87	359
Effective obs. right	435	422	417	457	87	378
Polyn. order	1	1	1	1	1	1
Bandwidth	0.251	0.239	0.234	0.267	0.232	0.207
Mean, left of threshold	-0.049	0.121	0.330	0.090	-0.064	-0.061

(b) France

Outcome	Index	Text similarity	Ideology		Complexity		Topics
	Full sample	Full sample	Full sample	Middle points	Full sample	Middle points	Full sample
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	0.275*	0.109	0.182	0.438**	0.127	-0.047	0.254
	(0.135)	(0.127)	(0.124)	(0.163)	(0.111)	(0.173)	(0.152)
Robust p-value	0.061	0.445	0.188	0.015	0.357	0.843	0.106
Observations left	688	688	688	312	688	172	688
Observations right	721	721	721	312	721	172	721
Effective obs. left	431	490	360	183	399	100	413
Effective obs. right	445	503	368	183	412	100	427
Polyn. order	1	1	1	1	1	1	1
Bandwidth	0.046	0.056	0.034	0.042	0.041	0.036	0.043
Mean, left of threshold	-0.116	-0.250	0.230	-0.047	0.001	-0.104	-0.020

Notes: We control for a dummy indicating if the candidate's party in the U.S., or political orientation in France, won the previous election. Other notes as in Tables 2 and 3.

Table D.7: General balance tests for ideology and complexity middle point races

(a) U.S.

Outcome	Complexity
	Middle points
	(1)
Treatment	0.059
	(0.041)
Robust p-value	0.246
Observations left	207
Observations right	207
Effective obs. left	79
Effective obs. right	79
Polyn. order	1
Bandwidth	0.208
Mean, left of threshold	0.415

(b) France

Outcome	Complexity	Ideology
	Middle points	Middle points
	(1)	(2)
Treatment	0.031	0.007
	(0.037)	(0.018)
Robust p-value	0.601	0.686
Observations left	172	312
Observations right	172	312
Effective obs. left	98	175
Effective obs. right	98	175
Polyn. order	1	1
Bandwidth	0.034	0.038
Mean, left of threshold	0.462	0.485

Notes: The sample is restricted to races in which the leader is initially in the middle of the opponent and the runner-up on the complexity scale (column 1, Panels a and b) or the ideology scale (column 2, Panel b), and in which we observe the primary election websites of both primary election contenders (Panel a) or the first-round manifestos of both the second- and third-ranked candidates in the first round (Panel b). In the U.S., the general balance test for middle point races along the ideology scale cannot be estimated due to the limited number of elections where a Republican (Democratic) candidate is to the left (right) of a Democratic (Republican) candidate before the primary (19 elections in total). Other notes as in Table D.3.

Table D.8: Overall convergence (second round with exactly two candidates)

Sample	Pooled (1)	U.S. (2)	France (3)
Treatment	0.358***	0.486**	0.315**
	(0.117)	(0.205)	(0.141)
Robust p-value	0.003	0.016	0.045
Observations left	991	324	667
Observations right	1128	427	701
Effective obs. left	566	138	394
Effective obs. right	605	163	408
Polyn. order	1	1	1
Bandwidth	0.586	0.221	0.042
Mean, left of threshold	-0.110	-0.122	-0.104

Notes: The sample is restricted to races where exactly two candidates, the leader and the qualified opponent, are present in the second round. Specifically, we exclude general elections where third-party candidates are present and where a primary election winner drops out before the general election, in the U.S.; and runoffs where more than two candidates qualify for the second round, as well as runoffs where only two candidates are qualified but one of them drops out of the race, in France. Other notes as in Table 2.

Table D.9: Convergence on different complexity measures

(a) U.S.

Outcome	Complexity index	Entropy	MATTR	Subordinates
	Middle points (1)	Middle points (2)	Middle points (3)	Middle points (4)
Treatment	0.455*	0.712*	0.362	0.748
	(0.250)	(0.407)	(0.205)	(0.546)
Robust p-value	0.090	0.100	0.116	0.235
Observations left	207	227	218	199
Observations right	207	227	218	199
Effective obs. left	87	103	119	92
Effective obs. right	87	103	119	92
Polyn. order	1	1	1	1
Bandwidth	0.236	0.252	0.275	0.251
Mean, left of threshold	-0.064	-0.196	-0.011	-0.106

(b) France

Outcome	Complexity index	Entropy	MATTR	Subordinates
	Middle points (1)	Middle points (2)	Middle points (3)	Middle points (4)
Treatment	-0.042	-0.089	0.075	-0.446
	(0.174)	(0.250)	(0.268)	(0.327)
Robust p-value	0.871	0.667	0.743	0.115
Observations left	172	195	184	196
Observations right	172	195	184	196
Effective obs. left	100	119	110	95
Effective obs. right	100	119	110	95
Polyn. order	1	1	1	1
Bandwidth	0.036	0.040	0.042	0.031
Mean, left of threshold	-0.104	0.085	0.062	0.052

Notes: The outcome is the change in overall complexity (column 1), the change in words complexity or entropy (column 2), the change in lexical diversity or MATTR (column 3), and the change is subordinate use (column 4). For each measure, the sample is restricted to races in which the leader's complexity measure is initially between the opponent and the runner-up's measures. Other notes as in Table 3.

Table D.10: Convergence to the more extreme opponent (McCrary density test)

Sample	U.S.		France	
	Extreme ideology (1)	Extreme complexity (2)	Extreme ideology (3)	Extreme complexity (4)
Density gap	0.215	0.020	-0.070	-1.433
	(0.392)	(0.309)	(2.838)	(2.920)
Robust p-value	0.584	0.948	0.980	0.624
Observations left	327	320	372	309
Observations right	300	307	230	293
Effective obs. left	181	217	251	215
Effective obs. right	148	223	152	173
Bandwidth	0.320	0.418	0.056	0.057

Notes: We report the running variable density difference at the threshold following the test proposed by McCrary (2008). We use the R implementation of the rddensity package (Cattaneo, Jansson and Ma, 2018) to estimate the density gap at the threshold and to compute p-values. Other notes as in Table 4.

Table D.11: Convergence to the more extreme opponent (general balance test)

Sample	U.S.		France	
	Extreme ideology (1)	Extreme complexity (2)	Extreme ideology (3)	Extreme complexity (4)
Treatment	0.002	-0.028	0.008	0.001
	(0.035)	(0.039)	(0.020)	(0.025)
Robust p-value	0.870	0.416	0.768	0.842
Observations left	327	320	372	309
Observations right	300	307	230	293
Effective obs. left	166	139	210	190
Effective obs. right	166	150	158	183
Polyn. order	1	1	1	1
Bandwidth	0.284	0.228	0.044	0.046
Mean, left of threshold	0.398	0.399	0.367	0.456

Notes: The outcome is the candidate's predicted treatment status based on observable characteristics listed in the text. Other notes as in Table 4.

Table D.12: Convergence to the incumbent or the candidate of a strong orientation

Sample	France: opponent is	
	Incumbent	From strong orientation
	(1)	(2)
Treatment	0.918*	0.343
	(0.541)	(0.250)
Robust p-value	0.070	0.134
Observations left	34	207
Observations right	90	269
Effective obs. left	19	115
Effective obs. right	28	123
Polyn. order	1	1
Bandwidth	0.023	0.030
Mean, left of threshold	-0.301	-0.211

Notes: The outcome is the change in overall similarity between the leader and the opponent between election rounds, defined as the average of the standardized changes in vectorized text similarity as well as similarity in ideology, complexity, and topic distribution. It is divided by its standard deviation. There is one observation per race and the running variable is defined as the difference in vote shares between the incumbent candidate and the other potential opponent (column 1) and between the candidate whose orientation received the most votes in the district in the previous election and the other potential opponent (column 2). It is positive in races where the qualified opponent is the incumbent (column 1) or when their orientation received the most votes in the previous election (column 2). We exclude races where the incumbent is not one of the two potential opponents (column 1) and races where the two potential opponents are from the same orientation as well as races where one of the opponents' orientation is missing or non-classified (column 2). Other notes as in Table 1.

Table D.13: Convergence to the incumbent or the candidate of a strong orientation (McCrary density test)

Sample	France: opponent is	
	Incumbent	From strong orientation
	(1)	(2)
Density gap	5.715	0.810
	(5.729)	(3.280)
Robust p-value	0.319	0.805
Observations left	34	207
Observations right	90	269
Effective obs. left	27	147
Effective obs. right	58	167
Bandwidth	0.059	0.047

Notes: We report the running variable density difference at the threshold following the test proposed by McCrary (2008). Other notes as in Tables in D.10 and D.12.

Table D.14: Convergence to the incumbent or the candidate of a strong orientation (general balance test)

Sample	France: opponent is	
	Incumbent	From strong orientation
	(1)	(2)
Treatment	-0.033	0.026
	(0.081)	(0.028)
Robust p-value	0.686	0.313
Observations left	34	207
Observations right	90	269
Effective obs. left	19	133
Effective obs. right	28	151
Polyn. order	1	1
Bandwidth	0.023	0.041
Mean, left of threshold	0.605	0.529

Notes: The outcome is the candidate's predicted treatment status based on observable characteristics listed in the text. Other notes as in Table D.12.

Table D.15: Heterogeneity of the convergence by leaders' characteristics

(a) U.S.

Sample	Left-wing (1)	Right-wing (2)	Re-runner (3)	First-timer (4)	Incumbent (5)	Challenger (6)
Treatment	0.353	0.427*	0.478**	0.167	0.507*	0.201
	(0.203)	(0.218)	(0.220)	(0.310)	(0.280)	(0.257)
Robust p-value	0.103	0.072	0.030	0.740	0.068	0.553
Observations left	417	407	347	260	275	332
Observations right	513	521	442	331	338	435
Effective obs. left	220	220	196	131	146	165
Effective obs. right	231	222	227	123	169	167
Polyn. order	1	1	1	1	1	1
Bandwidth	0.256	0.269	0.249	0.320	0.209	0.324
Mean, left of threshold	-0.026	-0.075	-0.087	0.043	-0.121	0.043

(b) France

Sample	Left-wing (1)	Right-wing (2)	Re-runner (3)	First-timer (4)	Incumbent (5)	Challenger (6)
Treatment	0.301	0.359	0.183	0.352	0.304	0.249
	(0.197)	(0.203)	(0.203)	(0.200)	(0.219)	(0.191)
Robust p-value	0.130	0.150	0.522	0.122	0.223	0.308
Observations left	343	350	334	328	292	370
Observations right	365	361	340	355	294	401
Effective obs. left	195	222	187	190	171	212
Effective obs. right	203	227	190	196	172	219
Polyn. order	1	1	1	1	1	1
Bandwidth	0.039	0.046	0.042	0.037	0.045	0.037
Mean, left of threshold	-0.101	-0.131	-0.263	0.052	-0.302	0.047

Notes: The outcome is the change in overall similarity to the opponent or runner-up between election rounds, defined as the average of the standardized changes in vectorized text similarity as well as similarity in ideology, complexity, and topic distribution. It is constructed separately and divided by its standard deviation within each country. The sample is further restricted to left-wing leaders in column 1 (defined as having an ideology score below the median in the first round), right-wing leaders in column 2 (defined as having an ideology score above the median in the first round), leaders who ran in the previous election in column 3, first-time runners in column 4, incumbents in column 5, and challengers in column 6. Other notes as in Table 1.

Table D.16: Heterogeneity of the convergence by leaders' characteristics (continued)

(a) U.S.

Sample	Primary high votes (1)	Primary low votes (2)	Party high votes (3)	Party low votes (4)
Treatment	0.488	0.529	0.514**	0.105
	(0.380)	(0.340)	(0.247)	(0.278)
Robust p-value	0.159	0.148	0.033	0.850
Observations left	219	210	307	257
Observations right	277	270	367	356
Effective obs. left	101	125	190	131
Effective obs. right	108	124	214	138
Polyn. order	1	1	1	1
Bandwidth	0.213	0.299	0.244	0.369
Mean, left of threshold	-0.015	0.055	-0.105	0.052

(b) France

Sample	1st round high votes (1)	1st round low votes (2)	Orientation high votes (3)	Orientation low votes (4)
Treatment	0.467**	0.316	0.273	0.278
	(0.189)	(0.217)	(0.206)	(0.215)
Robust p-value	0.020	0.139	0.179	0.289
Observations left	316	375	270	334
Observations right	339	384	290	335
Effective obs. left	182	206	163	189
Effective obs. right	187	205	175	182
Polyn. order	1	1	1	1
Bandwidth	0.051	0.027	0.049	0.032
Mean, left of threshold	-0.170	-0.076	-0.120	-0.103

Notes: The sample is restricted to leaders who received more than the median leader vote share in the first round in column 1, leaders who received less than the median vote share in column 2, leaders running in a district where their party (U.S.) or political orientation (France) received more than the median vote share during the previous general election in column 3, and leaders running in a district where their party or orientation received less than the median vote share in column 4. Other notes as in Table D.15.

Table D.17: Overall convergence in different time periods

Sample	U.S.		France	
	2002-2008	2010-2016	1978-2007	2012-2022
	(1)	(2)	(3)	(4)
Treatment	0.246	0.528***	0.233	0.352*
	(0.253)	(0.173)	(0.301)	(0.165)
Robust p-value	0.360	0.003	0.484	0.052
Observations left	282	539	222	466
Observations right	358	673	222	499
Effective obs. left	185	256	113	269
Effective obs. right	201	261	113	282
Polyn. order	1	1	1	1
Bandwidth	0.331	0.238	0.026	0.045
Mean, left of threshold	0.068	-0.110	-0.413	0.025

Notes: The sample is restricted to elections held until 2008 in columns 1 and 3, and to elections held after 2008 in columns 2 and 4. Other notes as in Table D.15.

E Additional Robustness Checks

E.1 Sample selection

United States As discussed in Section 5.1, our regression discontinuity sample for the U.S. includes elections in which there is, on the one hand, either a Republican or a Democratic competitive primary (i.e., a primary election with more than one candidate); and, on the other hand, a *leader* who is either a candidate of the opposite party or, in races in which there is no candidate of the opposite party, an independent candidate (if that candidate is the only other candidate). The qualification of a certain opponent against the closest contender in the primary election may potentially generate endogenous sample selection. For instance, the qualification of a more moderate Democratic candidate against an extreme one may discourage the Republican nominee and lead them to drop out of the race before the general election – in which case that particular race is not included in our sample. Below, we discuss the different ways in which our sample may be endogenously determined and how we address them:

- After a primary election, the defeated primary candidate may still receive votes in the general election. Between 2002 and 2016, this only happened in five elections. In all these cases, the defeated candidate had conceded. They were not formally listed on the general election ballot but received write-in votes, meaning that voters wrote their names on the ballot. Four of these five candidates received less than 0.05% votes and the fifth received 4.4% of votes. Since these candidates did not campaign for the general election and did not receive nearly enough votes to be threats, we do not expect the leader to adjust their behavior to them.
- The winner of a primary election may drop out before the general election. Between 2002 and 2016, this happened four times. Two candidates decided to retire for family reasons, one candidate had to drop out after being charged for domestic violence, and another gave her nomination to her husband. These candidates were replaced by other candidates from the same party. In three cases, the opposite party had run a competitive primary election, so it is possible that primary election winners took the identity of their general election opponent into account when they decided to drop out, which poses endogeneity concerns. However these cases are very rare, and our main result of overall convergence holds even when removing them: the point estimate in column 2 of Table E.1, Panel a, is virtually the same as the baseline estimate shown in column 1 (which reproduces the result shown in Table 2, column 2).
- The qualification of a more moderate primary candidate (e.g., Republican) may push the candidate from the opposite party (e.g., Democrat) not to run in the general election. If the opposite party itself ran a primary and the primary winner drops out before the general election, then this case would fall under the previous case. However, if the opposite party did not run a

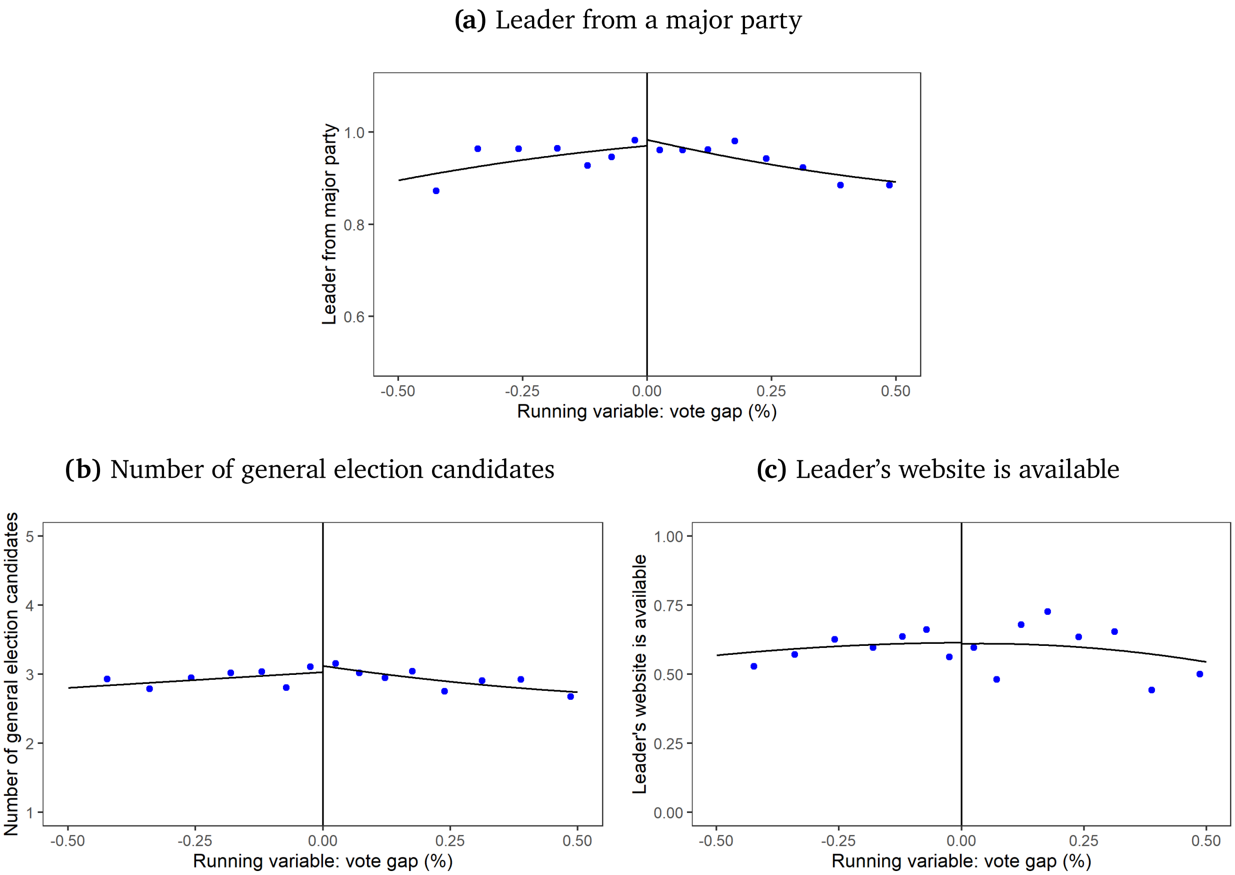
primary election, we are not able to determine whether the absence of a general election candidate from the opposite party was affected by the identity of the primary election winner on the other side – making our sample endogenously selected – or not. To assess the likelihood of this happening, we run a placebo test where the outcome is an indicator variable equal to one when the leader is endorsed by a major party (Republican or Democrat) and the running variable is the qualifying margin of the more extreme primary candidate from the opposite party (Figure E.1a). We also use the number of candidates running in the general election (Figure E.1b) and the availability of the leader's general election website on the Wayback Machine as alternative outcomes (Figure E.1c). We do not observe any jump at the threshold in any of these tests, suggesting that the leader being from a main party, the number of candidates running in a general election, and the availability of the leader's website, are not endogenously determined by the qualification of a more extreme opponent as opposed to a more moderate one. We also verify that our main result of overall convergence holds when restricting the sample to leaders running for a major party, Democrat or Republican (Table E.1, Panel a, column 3).

France In France, our sample includes elections in which the vote share of the candidate who ranked second in the first round is lower than the qualification threshold and in which the first-ranked candidate (the *leader*) competes in the runoff. In principle, the qualification of a certain opponent against the third-ranked candidate could endogenously affect the composition of our sample.

- A first concern is if the qualification of a more moderate opponent against an extreme one pushes the leader to drop out. However, in our RD sample, this situation never happens: all leaders are present in the second round.

- A second concern is if the qualification of an extreme opponent against a moderate candidate leads the qualified opponent to drop out, if they believe they stand no chance in the runoff. This would lead to the endogenous inclusion, in our sample, of leaders who should not respond to any convergence incentive since they do not have any opponent to converge to. There are 173 qualified second-ranked candidates who drop out before the runoff, accounting for 3.1% of all second-ranked candidates included in our RD sample. This issue also affects selection into the sample used in Appendix Table D.8, which includes races where both the leader and the qualified opponent are present in the runoff. We address this issue as follows. We note that the vast majority (75.7%) of the cases in which the qualified opponent drops out are races in which both the leader and the qualified opponent are left-wing. Therefore, excluding races in which the leader is left-wing and in which either the second-ranked or the third-ranked candidate is left-wing too enables us to exclude most races that could generate endogenous

Figure E.1: Placebo test: qualification of a more extreme candidate



(a) Leader from a major party

(b) Number of general election candidates

(c) Leader's website is available

Notes: In Figure E.1a, dots represent the local averages of an indicator variable equal to one when the leader is endorsed by a major party (Republican or Democrat). In Figure E.1b, dots represent the local averages of the number of general election candidates. In Figure E.1c, dots represent the local averages of an indicator variable equal to one when the leader's general election website is available on the Wayback Machine. Averages are calculated within quantile bins of the running variable. The running variable is the vote share difference between the most extreme candidate and their opponent in the primary election. It is measured in percentage points. The treatment variable is a dummy equal to 1 if the most extreme candidate qualifies for the general election. There is one observation per race and the sample is restricted to primary winners (or runner-ups) whose absolute ideological score before the primary is larger than that of their runner-up (winner). Continuous lines are a quadratic fit.

sample selection. The results of this robustness check are shown in column 2 of Table E.1, Panel b. We obtain a point estimate that is even larger in magnitude than the baseline estimate shown in column 1 (which reproduces the result shown in Table 2, column 3). If anything, this indicates that our main results may be slightly attenuated by the inclusion of leaders who have no opponent to converge to.

Table E.1: Overall convergence (robustness)

(a) U.S.

Sample	Baseline (1)	Robustness 1 (2)	Robustness 2 (3)
Treatment	0.414***	0.414***	0.420***
	(0.153)	(0.153)	(0.155)
Robust p-value	0.010	0.010	0.010
Observations left	821	820	812
Observations right	1031	1029	1016
Effective obs. left	425	429	418
Effective obs. right	439	443	432
Polyn. order	1	1	1
Bandwidth	0.253	0.255	0.251
Mean, left of threshold	-0.049	-0.049	-0.044

(b) France

Sample	Baseline (1)	Robustness (2)
Treatment	0.300**	0.378**
	(0.139)	(0.148)
Robust p-value	0.049	0.011
Observations left	688	645
Observations right	721	567
Effective obs. left	417	398
Effective obs. right	432	348
Polyn. order	1	1
Bandwidth	0.044	0.044
Mean, left of threshold	-0.116	-0.090

Notes: Column 1 of Panels a and b reports the baseline estimate of the overall convergence as in Table 2. In column 2, Panel a, the sample is restricted to races where the primary election winner is present in the general election. In column 3, Panel a, the sample only contains leaders associated with a major party (Democrat or Republican). In column 2, Panel b, the sample excludes elections in which the leader is left-wing and in which either the second- or third-ranked candidate is left-wing too. Other notes as in Table 1.