# The Shifting Attention of Political Leaders: Evidence from Two Centuries of Presidential Speeches*

Oscar Calvo-González Axel Eizmendi Germán Reyes

June 2023

**Abstract**

We use natural-language-processing algorithms on a novel dataset of over 900 presidential speeches from ten Latin American countries spanning two centuries to study the dynamics and determinants of presidential policy priorities. We show that most speech content can be characterized by a compact set of policy issues whose relative composition exhibited slow yet substantial shifts over 1819-2022. Presidential attention initially centered on military interventions and the development of state capacity. Attention gradually evolved towards building physical capital through investments in infrastructure and public services and finally turned towards building human capital through investments in education, health, and social safety nets. We characterize the way in which president-level characteristics, like age and gender, predict the main policy issues. Our findings offer novel insights into the dynamics of presidential attention and the factors that shape it, expanding our understanding of political agenda-setting.

*Calvo-González: The World Bank (ocalvogonzalez@worldbank.org), Eizmendi: Tufts University (axel.eizmendi_larrinaga@tufts.edu), Reyes: briq Institute on Behavior & Inequality (german.reyes@briq-institute.org). For helpful comments, we thank Aviv Caspi, Laura Chioda, Alexandru Cojocaru, Christa Deneault, Bill Maloney, Ambar Narayan, Roy Van der Weide, Daniel Valderrama, and participants from the Poverty GP lunch seminar. We also thank Evangelina Cabrera, Martha Narvaez, Sandra Pastrán, Cesar Polit, Yalile Uarac, for compiling and forwarding to us multiple speeches from Argentina, Chile, Ecuador, and Paraguay. This research was funded by the World Bank Research Department. Errors and omissions are our own.

## 1 Introduction

Issue attention—the policy issues that political actors pay attention to—is a key input to models of policymaking. For example, according to agenda-setting theories, issue attention is required to generate policy changes (e.g., Kingdon and Stano, 1984; Carmines and Stimson, 1986; Baumgartner and Jones, 2010). Similarly, the temporal patterns of issue attention are crucial for understanding the timing of future policy changes (Downs, 1972; Peters and Hogwood, 1985; Cairney, 2019).

Recent work uses natural-language-processing (NLP) algorithms to study political attention (Quinn et al., 2010; Grimmer and Stewart, 2013). By analyzing vast quantities of text, these methods can uncover the main issues discussed, under the assumption that issue attention is revealed by the relative allocation of expressed content. Automatized statistical algorithms have been used to study the content of treaties (Spirling, 2012), political e-mails (Mathur et al., 2023), legislators' tweets (Barbera et al., 2019), Federal Open Market Committee meetings (Hansen et al., 2018; Caspi and Stiglitz, 2020), and congressional speeches (e.g., Herzog and Benoit, 2015; Goet, 2019; Osnabrügge et al., 2021).

Yet, there has been little work using automatized methods to measure the political attention of presidents. The lack of such a study is significant given the central role of the presidency in shaping and directing policy. In this paper, we apply NLP methods to a novel dataset of presidential speeches to uncover the main expressed policy priorities of presidents, study their determinants, and analyze how they evolve over time.

We use a hand-collected dataset of over 900 annual presidential "state-of-the-union"-type speeches spanning ten Latin American countries. In these speeches, presidents provide an overview of policies undertaken by their administration and reflect on the priorities for the upcoming years. Our dataset dates as far back as 1819, enabling us to examine presidential discourse throughout significant historical periods. Among other events, our dataset covers a wide range of military conflicts, starting with the independence wars in which Latin American countries gained autonomy from Europe and covering both World Wars; multiple economic crises, including both the Great Depression and the Great Recession; and the rise to power of extremist leaders, both in the far right in the form of military dictatorships and the far left in the form of populist regimes.

To recover the policy issues discussed in the presidential speeches, we use a natural-language-processing algorithm called Latent Dirichlet Allocation (LDA). LDA uses the words in a set of documents as the only observable variables (Blei et al., 2003; Blei and

Lafferty, 2006; Blei, 2012). An attractive property of LDA is that it does not require the researcher to specify a set of topics into which the documents are classified.[1] LDA partitions the dataset of presidential speeches into a set of mutually exclusive and collectively exhaustive "topics." A topic is defined by a probability distribution over the keywords contained in the dataset of speeches. LDA also generates the probability distribution of topics of a given president's speech, which can be interpreted as the proportion of a president's speech discussing each topic. This measure is often used in the literature as a proxy for issue attention.

We use the topics uncovered by LDA for three purposes. First, we show that, despite the high dimensionality of presidential speeches, the expressed policy priorities embedded in the speeches can be characterized by a compact and easy-to-interpret set of issues. Most speech content falls into one of six topics: (i) military conflict and patriotism; (ii) the state of the public administration; (iii) investments in infrastructure; (iv) freedom, individual rights, and social justice; (v) economic development; and (vi) social protection. Across countries and years, these six policy issues together account, on average, for approximately 80% of presidential speech content.

Second, we study the president-level correlates of expressed presidential priorities. We find that female presidents, older presidents, and democratically-elected presidents are, on average, more likely to discuss economic development and social protection and less likely to discuss war/patriotism and the state of the public administration. Female leaders are also less likely to discuss issues related to infrastructure than male presidents. These findings are consistent with previous work relating political actors' traits to their expressed priorities (e.g., Gennaro and Ash, 2021; Osnabrügge et al., 2021) or their constituents' outcomes (e.g., Schubert, 1988; Chattopadhyay and Duflo, 2004; Clots-Figueras, 2012),

Third, we investigate the dynamics of presidential priorities. We find that the topics discussed in speeches slowly shift over long periods of time that stretch across electoral cycles. Over the course of two centuries, presidential attention gradually shifted from military interventions and the development of state capacity, to building physical capital through investments in infrastructure and public services, and finally to promoting human capital through investments in education, health, and social safety nets. These findings are consistent with issue attention theories positing that policy priorities often exhibit

[1] LDA is an increasingly-popular tool in economics. Researchers typically use it to recover latent types from high-dimensional text data. For example, LDA has been used to measure communication in deliberative bodies (Hansen et al., 2018), to study how CEO type affects firm performance (Bandiera et al., 2020), and to measure the value of firm amenities covered by collective bargaining agreements (Lagos, 2021).

continuity and long-term trajectories due to the enduring nature of societal challenges and institutional constraints (Baumgartner and Jones, 2010).

The rest of the paper proceeds as follows. Section 2 describes the data. Section 3 describes topic models and characterizes the issues discussed in presidential speeches. Section 4 studies the correlates and dynamics of speech content. Section 5 concludes.

## 2 Data

We assembled a novel dataset consisting of 933 presidential speeches delivered between 1819 and 2021 in ten Spanish-speaking Latin American countries: Argentina, Chile, Colombia, Costa Rica, The Dominican Republic, Ecuador, Mexico, Paraguay, Peru, and Venezuela. These constitutionally-required annual addresses serve as the closest parallel to the United States' "State of the Union" speech in these countries. In these speeches, presidents provide an overview of the work performed by their administration and an outline of the policy goals and priorities for the upcoming years.

The dataset compilation consisted of a two-stage process: collecting the speeches and processing them. We obtained the majority of speeches from Argentina, Ecuador, and Paraguay through their respective National Congresses and librarians. In the case of Chile, Colombia, Costa Rica, the Dominican Republic, and Mexico, speeches were collected through a variety of online sources. Most of Venezuela's speeches were scanned from books available at the US Library of Congress.

The second stage involved processing the speeches to enable text analysis. For the speeches already in a digitized format, this meant converting each file into text format and removing any text that did not form part of the speech, such as the title or date. For scanned speeches—i.e., those in "image" format—we performed Optical Character Recognition (OCR) to convert the images into machine-encoded text. To ensure the quality of the data, we manually reviewed all the OCR-generated text and corrected any inaccuracies. The text analysis was performed in Spanish. English translations shown throughout the paper were done by the authors.

Due to variability in data availability, both online and in the Congressional libraries, not every country nor every decade is represented equally in the dataset (see Appendix B). Costa Rica, Mexico, Peru, and Venezuela are the most represented countries in the dataset, each with 15–18% of all speeches. Argentina, Chile, Ecuador, and Paraguay have a moderate representation, each with 7%–10% of all speeches. Colombia and the Dominican

Republic are equally underrepresented in the dataset, each with only 2% of total speeches. The majority of speeches (68%) correspond to 1920–2021, with the remaining 32% dating back to 1819–1919.

We complement the presidential speeches with data on president demographic characteristics and political regime type. We obtain data on presidential terms and presidents' demographic characteristics (age and gender) from Archigos (Goemans et al., 2009). To classify governments as autocratic or democratic, we use data from the Polity 5 project (Marshall and Gurr, 2020). We define presidents as democratically elected if their "polity score" is positive (Persson and Tabellini, 2009).

Table 1 provides summary statistics on the dataset. Most presidents in our sample are males (97.6%). Presidents were, on average, 54.6 years old when they delivered their speeches. There is substantial variation in regime type in the dataset, with about half of the speeches corresponding to democratically-elected presidents (51.6%) and the other half to autocratic ones (48.4%). The party of almost half the presidents (45.8%) controlled both legislative chambers when the presidents delivered their speeches.

Table 1: Summary statistics on the sample of presidential speeches

| | Mean (1) | SD (2) | N (3) |
|---|---|---|---|
| **Panel A. Characteristics of the speeches** | | | |
| Delivered during 1819–1900 | 0.225 | 0.418 | 933 |
| Delivered during 1901–1950 | 0.268 | 0.443 | 933 |
| Delivered during 1951–2000 | 0.301 | 0.459 | 933 |
| Delivered during 2001–2021 | 0.206 | 0.404 | 933 |
| Words in speech | 5,637.7 | 5,323.6 | 933 |
| **Panel B. Characteristics of the presidents** | | | |
| Age | 54.8 | 9.6 | 780 |
| Male | 0.976 | 0.154 | 780 |
| Days in office | 2,776.2 | 2,925.1 | 780 |
| Democratically-elected | 0.516 | 0.500 | 902 |
| President's party controls both Houses | 0.458 | 0.499 | 306 |

*Notes:* This table shows summary statistics on our dataset. President characteristics come from Archigos and are typically available since 1870. Regime type (democracy/autocracy) comes from Polity 5 and is available since 1820. Party control of the Congress and the Senate come from the Database of Political Institutions (DPI) and is available since 1975.

# 3 The Expressed Policy Priorities in Presidential Speeches

This section describes topic models and examines the main policy issues discussed by presidents in our dataset of speeches.

## 3.1 Topic Models and Latent Dirichlet Allocation

Topic models are statistical models used to extract the main themes contained in large, unstructured collections of documents (Blei, 2012). The most widely used topic model is the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003). Given a collection of documents, LDA discovers the primary topics in each document and the degree to which each document exhibits those topics. LDA assumes that documents are random probability distributions over topics, where a topic is a probability distribution over the documents' words (see Appendix D.1). The only object that is exogenously defined by the researcher is the number of topics to be discovered. However, there are procedures to optimally select this figure, such as perplexity minimization (see Appendix D.4).

We use LDA to partition each presidential speech into a set of mutually exclusive and collectively exhaustive topics. To choose the number of topics, we follow the perplexity-minimization criterion, which yields 25 topics (Appendix Figure A1). For each topic, LDA produces a vector of keywords and the likelihood of each keyword belonging to the topic. We assign labels to the topics based on the most probable words to ease interpretability. LDA also generates the probability distribution of topics in each speech, which we interpret as the proportion of a speech discussing each topic.

## 3.2 The Policy Priorities Embedded in Presidential Speeches

Despite the complexity and high-dimensionality of presidential speeches, we find that a small set of policy issues can be used to characterize most speech content. There are only six topics that exceed 20% of speech content in at least one decade. Table 2 lists and defines these six topics. The labels we assign to these topics based on their most representative terms are: (i) War and patriotism, (ii) Public administration, (iii) Infrastructure, (iv) Rights, freedom, and social justice, (v) Economic development, and (vi) Social protection. On average, these six issues jointly cover 77.1% of speech content across countries and years. Appendix Table C1 shows that the main topics discovered by LDA are remarkably similar when varying the number of topics to be discovered by the algorithm.

Table 2: Top ten keywords defining the main topics of presidential speeches

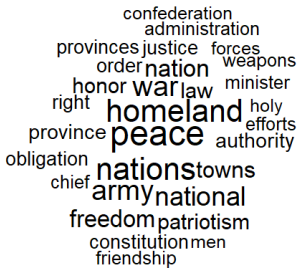
| | Topic name | | | | | |
|---|---|---|---|---|---|---|
| | War and Patriotism (1) | Public administration (2) | Infrastructure (3) | Rights and freedom (4) | Economic development (5) | Social protection (6) |
| 1 | Peace | Public | National | Town | Developing | Health |
| 2 | Towns | National | Plays | Politics | National | Family |
| 3 | Homeland | Service | Production | Social | Social | Education |
| 4 | War | Plays | Construction | Life | Program | Program |
| 5 | Army | Management | Education | Right | Politics | Social |
| 6 | National | Executive | Services | Liberty | Sector | Quality |
| 7 | Nation | Right | Social | National | System | Safety |
| 8 | Liberty | Interests | Plan | Homeland | Means | Poverty |
| 9 | Town | Public | Service | Nation | Process | Right |
| 10 | Law | Order | Activities | Economical | Increase | Investment |

*Notes*: This table lists the top ten keywords that define the topics in Figure 2. We only list topics whose probability exceeds 20% in at least one decade. Each column header shows the manually-assigned label of each topic. We recover the topics and their probability distributions using a Latent Dirichlet Allocation (LDA) algorithm (see Section 3.1 and Appendix D).

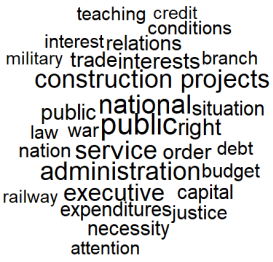
Figure 1 displays a series of word clouds plotting the top-defining keywords of each topic. Appendix Table A1 illustrates the content of each topic by providing excerpts from presidential speeches in which LDA assigns a high probability of belonging to a given topic. We discuss in more detail the content of each topic in the context of how the prevalence of these topics in presidential speeches evolved over time.

Figure 1: Word clouds of the keywords that define main topics

Panel A. War and patriotism

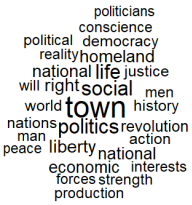



Panel B. Public administration




Panel C. Infrastructure


Panel D. Rights, freedom, and social justice




Panel E. Economic development


Panel F. Social protection


*Notes:* This figure plots word clouds of the distribution of words that define each topic. The size of each keyword is proportional to the importance of the keyword in defining the topic. We estimate the topics and their probability distributions using a Latent Dirichlet Allocation (LDA) algorithm (see Appendix D).

## 4 The Determinants and Dynamics of Policy Priorities

This section examines the correlates and evolution of the main policy issues discussed by presidents.

### 4.1 The President-level Correlates of Expressed Political Priorities

We begin by analyzing whether president characteristics correlate with expressed priorities. In Table 3, we estimate bivariate regressions of the form:

$$\text{ShareTopic}_{it}^{k} = \alpha^{k} + \gamma^{k} X_{it} + \varepsilon_{it}, \tag{1}$$

where $\text{ShareTopic}_{it}^{k}$ is the topic $k$ proportion in the presidential speech of country $i$ in year $t$ and $X_{it}$ is a president-level characteristic, such as age, gender, and political regime type (democracy/autocracy). We estimate equation (1) separately for each of the six main policy issues discussed in the presidential speeches.

President characteristics strongly correlate with expressed priorities (Table 3). Female presidents are 25.1 percentage points more likely to discuss social protection ($p < 0.01$), 6.0 percentage points more likely to discuss economic development ($p < 0.05$), 21.5 percentage points less likely to discuss public administration affairs ($p < 0.01$), 9.9 percentage points less likely to discuss infrastructure ($p < 0.01$), and 3.4 percentage points less likely to discuss war and patriotism ($p < 0.01$) than male presidents (columns 1, 2, 3, and 6).

Democratically-elected presidents are 16.7 percentage points more likely to discuss social protection ($p < 0.01$), 12.2 percentage points more likely to discuss economic development ($p < 0.01$), 17.1 less likely to discuss the state of the public administration, and 9.3 percentage points less likely to discuss war and patriotism ($p < 0.01$) than autocratic presidents (columns 1, 2, 5, and 6). President age follows the same qualitative patterns of expressed priorities as democratically-elected presidents.

In Appendix Table A2, we show that these patterns are quantitatively similar and qualitatively identical if all characteristics are simultaneously included in the regression equation.

Table 3: The president-level correlates of expressed policy priorities

| | Dependent variable: Fraction of a speech discussing... | | | | | |
|---|---|---|---|---|---|---|
| | War and Patriotism (1) | Public administration (2) | Infrastructure (3) | Rights and freedom (4) | Economic development (5) | Social protection (6) |
| Democracy | −0.093*** | −0.171*** | 0.012 | 0.013 | 0.122*** | 0.167*** |
| | (0.008) | (0.014) | (0.009) | (0.008) | (0.010) | (0.010) |
| Age | −0.001*** | −0.005*** | 0.001 | 0.000 | 0.003*** | 0.002*** |
| | (0.000) | (0.001) | (0.000) | (0.000) | (0.001) | (0.001) |
| Female | −0.034*** | −0.215*** | −0.099*** | −0.016 | 0.060*** | 0.251*** |
| | (0.004) | (0.009) | (0.009) | (0.038) | (0.023) | (0.046) |
| $N$ (speeches) | 780 | 780 | 780 | 780 | 780 | 780 |
| Mean DV | 0.039 | 0.228 | 0.143 | 0.106 | 0.147 | 0.096 |

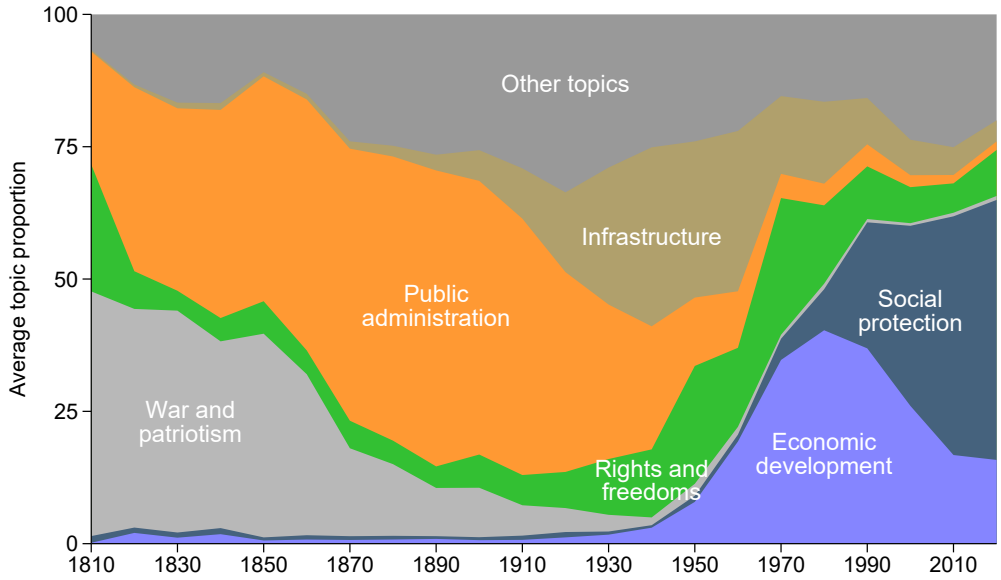
*Notes:* This table presents OLS coefficients, $\gamma^k$, from equation (1) Each cell displays coefficients from a bivariate regression of the variable listed in the column header on the variable listed in the row header. Heteroskedasticity-robust standard errors in parentheses. ***, **, and * denote significance at the 1%, 5% and 10% levels, respectively.

The positive relationship between female leadership and content devoted to social protection—a topic that stresses the importance of investments in education and health—is consistent with the findings of Clots-Figueras (2012), who shows that female politicians increase educational attainment in urban areas in India. The finding that democratic regimes are more likely to discuss economic development is consistent with the literature that links democracy to economic growth (e.g., Acemoglu et al., 2019). Similarly, the link between democratic leadership and expressed attention to social protection is in line with findings showing a positive link between democratic regimes and health outcomes, such as life expectancy (e.g., Besley and Kudamatsu, 2006).

### 4.2 The Evolution of Expressed Priorities Over Time

Next, we turn to the dynamics of expressed priorities. Figure 2 plots the distribution of the main issues discussed in presidential speeches in each decade from 1819–2021. To construct this figure, we aggregate the topics discussed in speeches at the decade level and calculate the average topic proportion in each decade. This enables us to focus on the main topics discussed in each decade but at the cost of ignoring the idiosyncratic year-to-year variation in presidential discourse (captured in our residual "Other Topics" category).

Figure 2: The evolution of expressed presidential priorities over time




*Notes:* This figure shows the distribution of topics across decades. We estimate the topics and their probability distributions using a Latent Dirichlet Allocation (LDA) algorithm (see Section 3.1 and Appendix D). To choose the number of topics, we follow a criterion of perplexity minimization (see Appendix D.4). We present only topics whose probability exceeds 20% in at least one decade. The rest of the topics are grouped together in the category labeled as "Other topics." Topics are defined by their top occurring keywords (see Table 2 for the top ten keywords that define the topics in the figure). We manually labeled topics based on the top keywords. To construct the figure, we pool speeches at the decade level and compute the average topic probabilities in each decade.

The main topics discussed by presidents tend to change slowly over multiple decades that stretch across electoral cycles. Throughout the 19th century, presidents were primarily concerned with two topics: war/patriotism and the state of the public administration. These two topics combined accounted for roughly 70% of presidential speech content until the 1870s. The prevalence of the war and patriotism topic is consistent with the historical context of the period, which was characterized by internal conflicts and repeated border wars between the newborn nation-states of Latin America (Clayton et al., 2017). The public administration topic captures a range of governance issues, including the organization of government establishments, the provision of public services, and the management of public finances. This likely reflects the fact that the governments and their institutions were still in their infancy, making the development of political order and state capacity a central policy concern (Clayton et al., 2017).

Discussions of war and patriotism decreased after the 1870s. Although the topic of pub-

lic administration continued to dominate well into the 20th century, beginning in the 1900s, presidents increasingly shifted their rhetoric towards the development of infrastructure and the provision of public services. These policy concerns were particularly dominant during 1920–1950, with attention to this topic reaching its peak in the 1940s. This is consistent with the growing disillusionment with export-led growth during the period and a subsequent shift towards import-substitution industrialization strategies, which emphasized the growth of internal markets, the development of infrastructure, and greater government intervention (Prebisch, 1962; Bulmer-Thomas, 2003). Around this time, presidents increasingly began to discuss a broad topic related to rights, freedom, and social justice, particularly from 1940–1980. Discussions about these issues are perhaps expected, given that this historical period was characterized by the increasing popularity of communist and leftist ideologies—as evidenced by the rise of revolutionary movements across the region—and the rise and fall of military dictatorships, often running on anti-communist agendas (Dávila, 2013).

Beginning in the 1950s, there is a stark increase in attention to economic development. During this period, presidents framed economic development as a byproduct of economic growth and changes in a country's productive structure. Accordingly, this topic captures discussions related to economic planning and the management of industrial sectors, which is consistent with the increased adoption of import-substitution industrialization development strategies (Bulmer-Thomas, 2003).

In the 1990s—a period characterized by the implementation of so-called Washington Consensus policies in several countries of the region (Williamson, 1993; Gore, 2000) and a wave of democratization (Hagopian and Mainwaring, 2005)—presidential discourse became increasingly concentrated on issues related to social protection. As the top keywords illustrate, this topic focuses on the importance of building human capital through investments in education and health and stresses the state's role in providing social insurance and building social safety nets. The appearance of the terms "right," "investment," and "poverty," suggests an increasing recognition that access to these public services constitutes a fundamental right that requires government investment and that education and health care are pathways to reducing poverty.[2] The high prevalence of this topic continued in the 2000s, a period characterized by the surge of leftist governments across Latin America.[3]

[2]Another top keyword of this topic is "programs." This is consistent with the rise of large-scale social programs throughout the region, such as conditional cash transfers and non-contributory pensions (Ferreira and Robalino, 2010).

[3]This movement, which is often considered to have started with Chavez's accession to power in Venezuela

During 2000–2021, social protection accrued, on average, over 40% of speech content and remains the dominant policy issue discussed in presidential speeches to this day.

These patterns reveal how presidents have characterized a country's priorities over time, shifting from military interventions and the development of state capacity, to building physical capital through investments in infrastructure and public services, and finally to promoting human capital through investments in education, health, and social safety nets. Moreover, while idiosyncratic policy issues comprise a non-trivial share of speech content (as reflected in a 5%–25% speech content share of the "Other Topics" residual category), we also find that a small number of key policy issues make up a significant portion of the priorities expressed by presidents.

In Appendix C, we conduct a series of robustness checks. First, we vary the number of LDA topics and show that the patterns are similar across the number of topics. Second, we divide our sample into four sub-periods and estimate LDA separately for each period. We find that the list of topics and their evolution over time are remarkably consistent with our baseline results.

## 5 Conclusion

In this paper, we combine data from presidential speeches with natural-language-processing algorithms to measure the expressed priorities of presidents. We show that high-dimensional speeches can be characterized by a compact set of policy issues. Consistent with political attention theories positing that issue attention remains stable for long periods (e.g., Baumgartner and Jones, 2010), we find that expressed presidential priorities shift slowly over time and span multiple electoral cycles. Moreover, in line with previous work relating political actors' traits to their policy preferences (e.g., Clots-Figueras, 2012; Besley and Kudamatsu, 2006), we find that presidential traits, such as age and gender, are correlated with issue attention. These novel empirical facts offer insights for political agenda-setting theories and expand our understanding of presidential issue attention.

Future research could leverage the dataset of presidential speeches to measure whether the policy prioritization of an issue in a given country has spillover effects on the likelihood of other countries discussing the same issue in subsequent years (Buera et al., 2011). In addition, our dataset can be used to learn how the conceptualization of different policy

at the beginning of 1999, is considered to include the elections of Evo Morales in Bolivia, Rafael Correa in Ecuador, and Daniel Ortega in Nicaragua, all in 2006, and to a lesser extent Luiz Ignacio Lula da Silva in Brazil in 2002, Nestor Kirchner in Argentina in 2003, and Tabaré Vázquez in Uruguay in 2005.

tools (such as personal income taxes or environmental regulations) changes over time and how these correlate with measures of actual policy implementation.

## References


Acemoglu, D., S. Naidu, P. Restrepo, and J. A. Robinson (2019). Democracy does cause growth. *Journal of political economy 127*(1), 47–100.

Bandiera, O., A. Prat, S. Hansen, and R. Sadun (2020). Ceo behavior and firm performance. *Journal of Political Economy 128*(4), 1325–1369.

Barbera, P., A. Casas, J. Nagler, P. J. Egan, R. Bonneau, J. T. Jost, and J. A. Tucker (2019). Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review 113*(4), 883–901.

Baumgartner, F. R. and B. D. Jones (2010). *Agendas and instability in American politics*. University of Chicago Press.

Besley, T. and M. Kudamatsu (2006). Health and democracy. *American economic review 96*(2), 313–318.

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

Blei, D. M. and J. D. Lafferty (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120.

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

Buera, F. J., A. Monge-Naranjo, and G. E. Primiceri (2011). Learning the wealth of nations. *Econometrica 79*(1), 1–45.

Bulmer-Thomas, V. (2003). *The economic history of Latin America since independence*. Cambridge University Press.

Cairney, P. (2019). *Understanding public policy: theories and issues*. Bloomsbury Publishing.

Carmines, E. G. and J. A. Stimson (1986). On the structure and sequence of issue evolution. *American Political Science Review 80*(3), 901–920.

Caspi, A. and E. H. Stiglitz (2020). Measuring discourse by algorithm. *International Review of Law and Economics 62*, 105863.

Chattopadhyay, R. and E. Duflo (2004). Women as policy makers: Evidence from a randomized policy experiment in india. *Econometrica 72*(5), 1409–1443.

Clayton, L. A., M. L. Conniff, and S. M. Gauss (2017). *A New History of Modern Latin America* (3 ed.). University of California Press.

Clots-Figueras, I. (2012). Are female leaders good for education? evidence from india. *American Economic Journal: Applied Economics 4*(1), 212–44.

Dávila, J. (2013). *Dictatorship in South America*. John Wiley & Sons.

Downs, A. (1972). Up and down with ecology: The issue-attention cycle. *The public 28*, 38–50.

Ferreira, F. H. and D. A. Robalino (2010). Social protection in latin america: achievements and limitations. *World Bank Policy Research Working Paper* (5305).

Gennaro, G. and E. Ash (2021, 12). Emotion and Reason in Political Language. *The Economic Journal 132*(643), 1037–1059.

Goemans, H. E., K. S. Gleditsch, and G. Chiozza (2009). Introducing archigos: A dataset of political leaders. *Journal of Peace research 46*(2), 269–283.

Goet, N. D. (2019). Measuring polarization with text analysis: evidence from the uk house of commons, 1811–2015. *Political Analysis 27*(4), 518–539.

Gore, C. (2000). The rise and fall of the washington consensus as a paradigm for developing countries. *World development 28*(5), 789–804.

Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National academy of Sciences 101*(suppl 1), 5228–5235.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis 21*(3), 267–297.

Grün, B. and K. Hornik (2011). topicmodels: An r package for fitting topic models. *Journal of statistical software 40*, 1–30.

Hagopian, F. and S. P. Mainwaring (2005). *The third wave of democratization in Latin America: advances and setbacks*. Cambridge University Press.

Hansen, S., M. McMahon, and A. Prat (2018). Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics 133*(2), 801–870.

Herzog, A. and K. Benoit (2015). The most unkindest cuts: speaker selection and expressed government dissent during economic crisis. *The Journal of Politics 77*(4), 1157–1175.

Kim, D. and A. Oh (2011). Topic chains for understanding a news corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 163–176. Springer.

Kingdon, J. W. and E. Stano (1984). *Agendas, alternatives, and public policies*, Volume 45. Little, Brown Boston.

Lagos, L. (2021). Labor market institutions and the composition of firm compensation: Evidence from brazilian collective bargaining. Working Paper.

Marshall, M. G. and T. R. Gurr (2020). Polity5: Political regime characteristics and transitions, 1800-2018. Dataset Users' Manual. Center for Systemic Peace. Available at http://www.systemicpeace.org/inscr/p5manualv2018.pdf.

Mathur, A., A. Wang, C. Schwemmer, M. Hamin, B. M. Stewart, and A. Narayanan (2023). Manipulative tactics are the norm in political emails: Evidence from 300k emails from the 2020 us election cycle. *Big Data & Society 10*(1), 20539517221145371.

Osnabrügge, M., E. Ash, and M. Morelli (2021). Cross-domain topic classification for political texts. *Political Analysis*, 1–22.

Osnabrügge, M., S. B. Hobolt, and T. Rodon (2021). Playing to the gallery: Emotive rhetoric in parliaments. *American Political Science Review 115*(3), 885–899.

Persson, T. and G. Tabellini (2009). Democratic capital: the nexus of economic and political change. *American Economic Journal Macroeconomics 1*, 88–126.

Peters, B. G. and B. W. Hogwood (1985). In search of the issue-attention cycle. *The journal of politics 47*(1), 238–253.

Prebisch, R. (1962). The economic development of latin america and its principal problems. *Economic Bulletin for Latin America*.

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

Schubert, J. N. (1988). Age and active-passive leadership style. *American Political Science Review 82*(3), 763–772.

Spirling, A. (2012). Us treaty making with american indians: Institutional change and relative power, 1784–1911. *American Journal of Political Science 56*(1), 84–97.

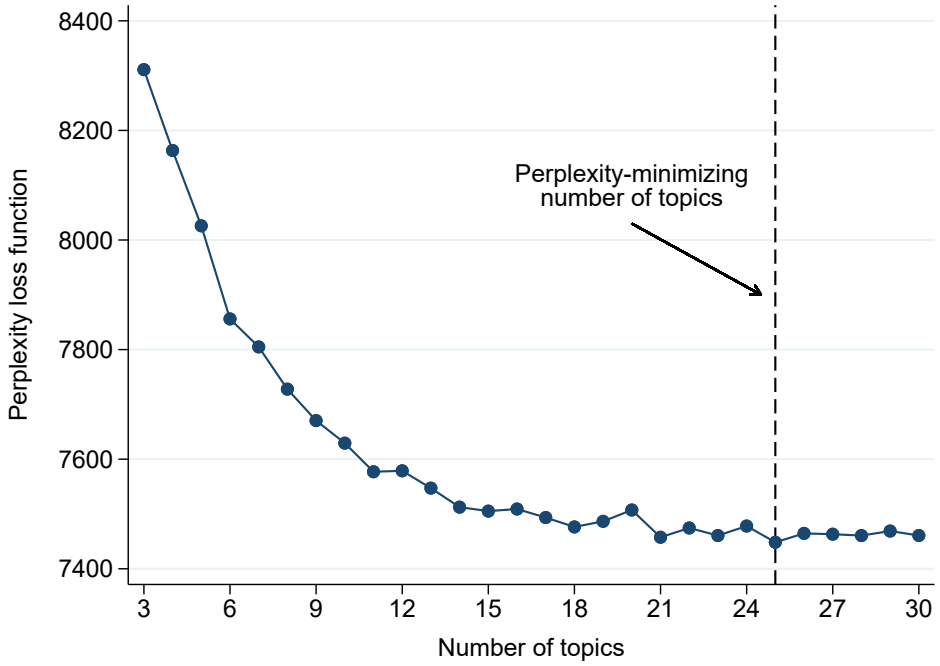
Williamson, J. (1993). Democracy and the "washington consensus". *World development 21*(8), 1329–1336.

Yildirim, I. (2012). Bayesian inference: Gibbs sampling. *Technical Note, University of Rochester*.

# Appendix — For Online Publication

## A Additional Figures and Tables

Figure A1: Perplexity and number of topics discussed in presidential speeches




*Notes:* This figure shows the estimated perplexity loss function against the number of topics in our sample of presidential speeches. To construct this figure, we follow a three-step process: (i) randomly partition our sample into a training set (90% of speeches) and a hold-out test set (10% of speeches), (ii) iteratively implement LDA on the training set using a different number of topics in each iteration, and (iii) compute the perplexity of each model. See Appendix D.4 for details about perplexity.

Table A1: Exercepts from presidential speeches illustrating the content of each topic

| Speech | Topic Prob. | Excerpt |
|---|---|---|
| (1) | (2) | (3) |
| **Panel A. War and Patriotism** | | |
| Peru, 1835 | 0.63 | The triumph of the United Army in Yanacocha, the new order following this battle, the performance of the United Army, the fixing of the national treasury, the peace in the towns and respect towards people and property, are testimony in favor of the government and against its detractors. |
| Venezuela, 1850 | 0.57 | Tireless and tenacious, the enemies of Venezuela have made formidable efforts to bring down the institutions and introduce a system that is diametrically opposed to the nation's will. [...] The government, always behaving frankly and generously, never withdrew its merciful hand and attempted to attract the strayed to the heart of the community in order to extirpate any germ of new revolts and rebellions, and cement in this way the peace that the towns need so much. |
| **Panel B. Public Administration** | | |
| Costa Rica, 1897 | 0.78 | The practice of removing Judges or Mayors only if there is a judicial decision to imprison them is another obstacle, perhaps larger in magnitude, that hinders good public service. [...] this reform would enable more effective and expedited action in the internal order of the Judiciary and would stimulate lower-ranking public servants to strictly comply with the delicate mission they have been entrusted with. |
| Peru, 1895 | 0.75 | The task of investigating the main administrative branches required staff exclusively dedicated to it, and to satisfy this important requirement, in addition to the inspectors of the Treasury and Customs, three investigative commissions were created: one tasked with examining the fiscal contracts of the previous administration; another with investigating public expenditures; and the third tasked with inspecting the state of the Callao Customs and the cause of its recent weakness. |
| **Panel C. Infrastructure** | | |
| Peru, 1959 | 0.53 | During the third year of my administration, 1,379km of roads have been constructed, with an investment of 234 million soles, and work has been done to maintain 12,500km of roads at a cost of more than 60 million soles... These construction projects will allow the promotion and introduction of basic industries, and will open new work centers. |
| Argentina, 1940 | 0.49 | Concerning the public works, the Executive has proposed to carry out actions oriented towards establishing a uniform type of construction within a model that satisfies the technical requirements of its intended use, procuring the suppression of monumental buildings... |

Table A1: Exercepts from presidential speeches illustrating the content of each topic (continued)

| Speech | Topic Prob. | Excerpt |
|---|---|---|
| (1) | (2) | (3) |
| **Panel D. Rights and Freedom** | | |
| Argentina, 1953 | 0.69 | In order to establish political sovereignty, I gave every Argentinian individual freedom in the effective enjoyment of all their rights, which arise from the dignity that can only be enjoyed by men who have been economically liberated by social justice... |
| Peru, 1973 | 0.62 | This means recognizing the right of others to think and act differently from us, and consequently, to organize themselves politically with complete freedom within a plurality of alternatives. Our Revolution represents one of these alternatives. |
| **Panel E. Economic Development** | | |
| Mexico, 1983 | 0.70 | We have presented and initiated actions to induce qualitative changes to the economic structure, to revise attitudes and update styles in order to improve the orientation and quality of development, and to transform it into a steady and sustained process |
| Venezuela, 1989 | 0.58 | We are currently on the path towards economic growth and social progress, so that the decade of the 90s becomes the decade of development. |
| **Panel F. Social Protection** | | |
| Mexico, 2008 | 0.72 | I have no doubt that we will continue to work intensely to build a fairer future for you and your family; to build a more humane Mexico, a Mexico with sustainable human development; a Mexico without extreme poverty; a Mexico with health and education for all. |
| Chile, 2012 | 0.61 | Education is the key engine of development and social mobility. It is the mechanism needed for the talent and merit to emerge. It is the great instrument for the construction of a country of opportunities. For this reason, the battle for development and against poverty will be won or lost in the classrooms. |

*Notes:* This table shows speech excerpts that illustrate the six main topics identified by LDA. We show excerpts from speeches with a high topic probability. We estimate the topics and their probability distributions using a Latent Dirichlet Allocation (LDA) algorithm (see Section 3.1 and Appendix D). To choose the number of topics, we follow a criterion of perplexity minimization (see Appendix D.4). We present only topics whose probability exceeds 20% in at least one decade. Topics are defined by their top occurring keywords (see Table 2 for the top ten keywords that define the topics in the figure). We manually labeled topics based on the top keywords.

Table A2: Robustness of the president-level correlates of expressed priorities to alternative specifications

| | Dependent variable: Fraction of a speech discussing... | | | | | |
|---|---|---|---|---|---|---|
| | War and Patriotism (1) | Public administration (2) | Infrastructure (3) | Rights and freedom (4) | Economic development (5) | Social protection (6) |
| Democracy | −0.040*** | −0.125*** | −0.003 | 0.007 | 0.109*** | 0.132*** |
| | (0.005) | (0.015) | (0.010) | (0.009) | (0.012) | (0.009) |
| Age | −0.001*** | −0.004*** | 0.001 | 0.000 | 0.002*** | 0.001*** |
| | (0.000) | (0.001) | (0.000) | (0.000) | (0.001) | (0.001) |
| Female | −0.015*** | −0.155*** | −0.098*** | −0.019 | 0.009 | 0.190*** |
| | (0.003) | (0.011) | (0.010) | (0.039) | (0.024) | (0.046) |
| $N$ (speeches) | 780 | 780 | 780 | 780 | 780 | 780 |
| Mean DV | 0.039 | 0.228 | 0.143 | 0.106 | 0.147 | 0.096 |

*Notes:* This table is analogous to Table 3, but all variables enter simultaneously in the regression equation. Heteroskedasticity-robust standard errors in parentheses. ***, **, and * denote significance at the 1%, 5% and 10% levels, respectively.

# B Data Appendix

## B.1 Presidential Speeches Inclusion Criteria

We included countries into our sample based on two criteria. First, we restricted our search to Spanish-speaking countries in order to be able to pool speeches for text analysis. Second, we focused only on countries that had an annual, constitutionally-mandated presidential speech in which the president gives an overview of the work the government has performed in each legislative session, as well as an outline of the policy goals and priorities for the future. Appendix Table B1 lists the Spanish-speaking Latin American countries that have this constitutional mandate, as well as the specific articles in the constitution that establish this requirement. Our sample is composed of the subset of countries for which we could locate speeches across at least two decades.

Table B1: Countries with constitutionally-mandated presidential speeches

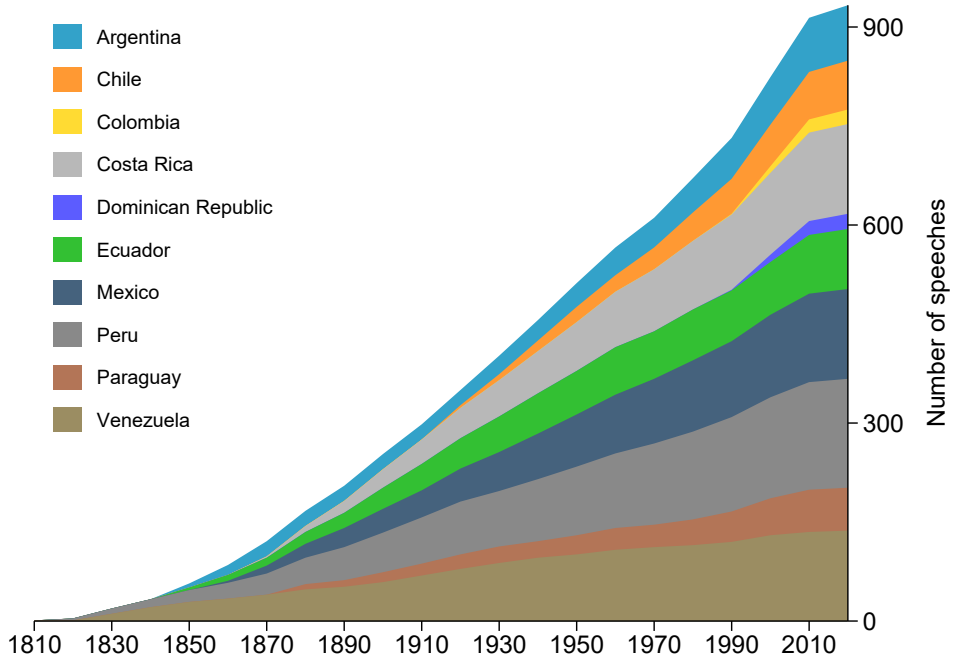
| Country | Article |
|---|---|
| (1) | (2) |
| Argentina | 99, subsection 8 |
| Bolivia | 96, subsection 10 |
| Colombia | 189, subsection 12 |
| Costa Rica | 139, subsection 4 |
| Chile | 24 |
| Ecuador | 171, subsection 7 |
| Guatemala | 183, subsection i |
| Honduras | 183, subsection i |
| Mexico | 69 |
| Nicaragua | 150, subsection 15 |
| Panama | 178, subsection 5 |
| Paraguay | 238, subsection 8 |
| Peru | 118, subsection 7 |
| Venezuela | 237 |
| Dominican Republic | 55, subsection 22 |
| Uruguay | 168, subsection 5 |

*Notes:* This table lists the Spanish-speaking Latin American countries with a constitutionally-mandated state-of-the-union-style speech and the article in the constitution that establishes this requirement.

## B.2 About the Missing Speeches

Appendix Figure B1 shows the cumulative sum of speeches by year.

Figure B1: Cumulative number of presidential speeches by country over time




*Notes:* This figure shows the cumulative number of speeches in the countries in our sample. See Appendix B for details on the data.

Although we made efforts to compile as many presidential speeches as possible, there are a number of countries for which the speeches from certain years are missing. In some years, presidential speeches are missing because no speech was delivered. For example, in the case of Paraguay, the president did not deliver a speech from 1940–1948. Political turmoil, such as a coup or an ongoing revolution, is another probable cause, particularly for missing speeches from the 1970s. Given the multiple coups and revolts in the region during this time, some of the missing speeches are likely associated with this turbulent political context.

It is hard to establish a cause with certainty for other missing speeches. In some cases, the congressional libraries could not locate the physical copies; in other cases, the quality of the original manuscripts was too low for proper digitalization. At times, particularly in the case of Venezuela, we could not find publications containing the compilations of the presidents' speeches for specific years.

# C Empirical Appendix

We assess the robustness of the Section 4.2 results in two ways. First, we vary the number of topics to be discovered by the LDA algorithm. We repeat our analysis using 5, 15, and 45 topics (Appendix Table C1 and Appendix Figure C1). The main topics uncovered by LDA are very similar when varying the number of topics and the evolution of these topics over time mirrors the one in the baseline analysis.

The second robustness check relates to a shortcoming of how LDA operates with time-series data. LDA assumes that the vocabulary is fixed over time (Blei and Lafferty, 2006). However, a significant change in language could affect topic discovery and assignment. To deal with this, we follow Kim and Oh (2011) and partition our sample into different periods. Then, we estimate the topics separately for each period. This allows LDA to discover the topics that dominated each period using only the words observed in that period instead of the words found in the entire vocabulary in our sample.

We divide our sample into four periods (with a similar number of observations per period): 1819–1900, 1901–1950, 1951–2000, and 2001–2021, and estimate the LDA algorithm on the speeches of each subperiod. Appendix Table C2 shows the main topics in each period and Appendix Figure C2 juxtaposes the evolution of these topics across periods. The list of topics is very similar to the list LDA yields when pooling all the periods together.

Table C1: Top ten keywords defining main topics

**Panel A. LDA using 5 topics**

| | Topic name | | | | |
|---|---|---|---|---|---|
| | War and Public administration (1) | Infrastructure (2) | Nationalism (3) | Social protection (4) | Economic development (5) |
| 1 | Executive | Plays | Village | Program | National |
| 2 | Public | National | World | Developing | Social |
| 3 | Nation | Service | National | Social | Politics |
| 4 | National | Construction | Homeland | National | Developing |
| 5 | Right | Services | Social | Health | Production |
| 6 | Administration | Department | Revolution | Sector | Village |
| 7 | Plays | Works | History | Education | Economical |
| 8 | Peace | School | Plan | System | Life |
| 9 | Order | Schools | Life | Increase | Program |
| 10 | Towns | Federal | Countries | Investment | Means |

**Panel B. LDA using 15 topics**

| | Topic name | | | | | |
|---|---|---|---|---|---|---|
| | War and Patriotism (1) | Public administration (2) | Infrastructure (3) | Rights and freedom (4) | Economic development (5) | Social protection (6) |
| 1 | Peace | Public | National | Village | Developing | Health |
| 2 | Towns | Administration | Plays | Politics | National | Family |
| 3 | Homeland | Service | Production | Social | Social | Education |
| 4 | National | Plays | Social | Life | Program | Quality |
| 5 | War | National | Services | Liberty | Sector | Poverty |
| 6 | Nation | Executive | Construction | Right | Politics | Draft |
| 7 | Army | Right | Activities | National | System | Security |
| 8 | Law | Interests | Education | Economical | Means | Right |
| 9 | Liberty | Order | Plan | Homeland | Process | Investment |
| 10 | Patriotism | Instruction | Service | Democracy | Increase | Woman |

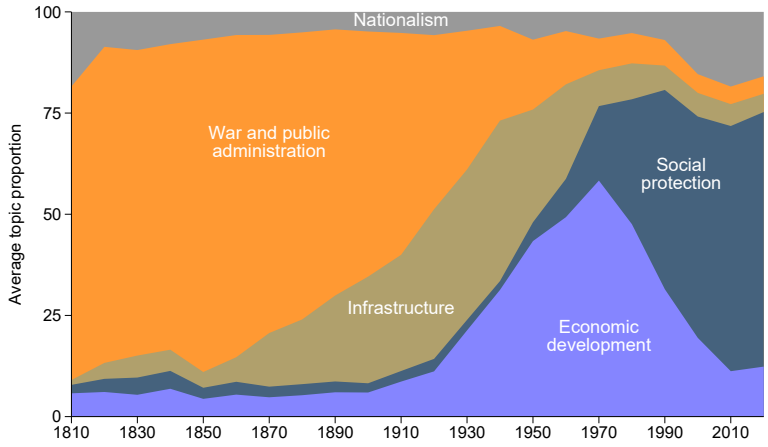
**Panel C. LDA using 45 topics**

| | Topic name | | | | | |
|---|---|---|---|---|---|---|
| | War and Patriotism (1) | Public administration (2) | Infrastructure (3) | Rights and freedom (4) | Economic development (5) | Social protection (6) |
| 1 | Peace | Public | National | Village | Developing | Health |
| 2 | Homeland | Plays | Plays | Politics | National | Education |
| 3 | Towns | National | Construction | Social | Social | Family |
| 4 | Army | Service | Production | Life | Program | Program |
| 5 | War | Administration | Services | Right | Politics | Investment |
| 6 | Nation | Executive | Education | Liberty | Sector | Quality |
| 7 | National | Right | Social | National | System | Security |
| 8 | Liberty | Public | Service | Homeland | Means | Right |
| 9 | Province | Interests | Plan25 | Nation | Process | Poverty |
| 10 | Honor | Relations | Ministry | Economical | Increase | Social |

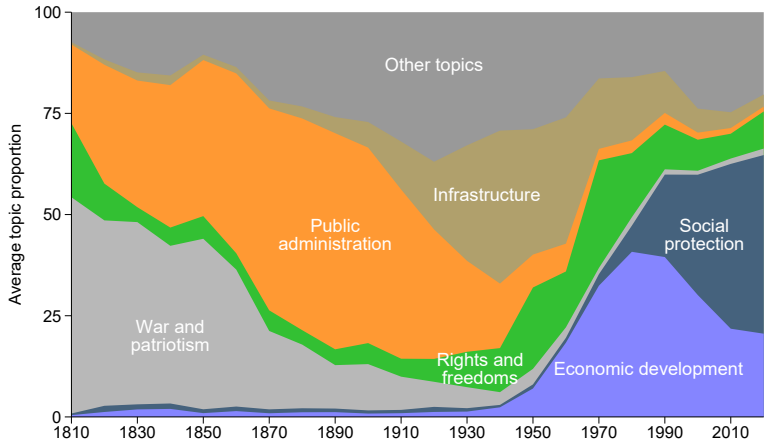
*Notes*: This table lists the top ten keywords that define the topics in Figure C1. We only list topics whose probability exceeds 20% in at least one decade. Each column header shows the manually-assigned label of each topic. We recover the topics and their probability distributions using a Latent Dirichlet Allocation (LDA) algorithm (see Section 3.1 and Appendix D).

Figure C1: Robustness of topics discovered by LDA to the number of topics in the speeches

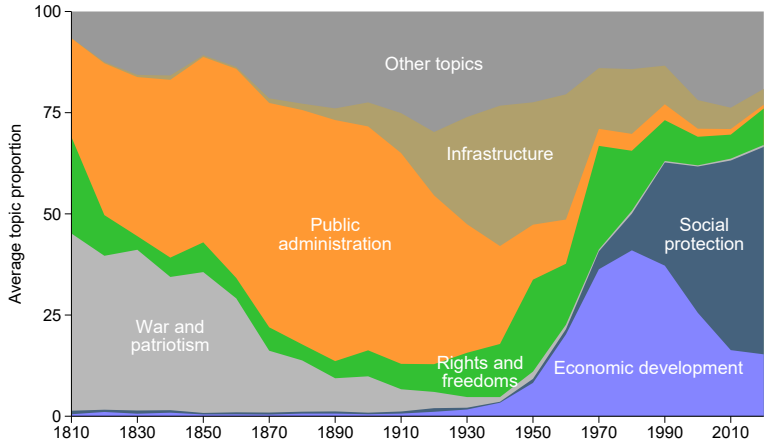
Panel A. Number of topics = 5




Panel B. Number of topics = 15




Panel C. Number of topics = 45





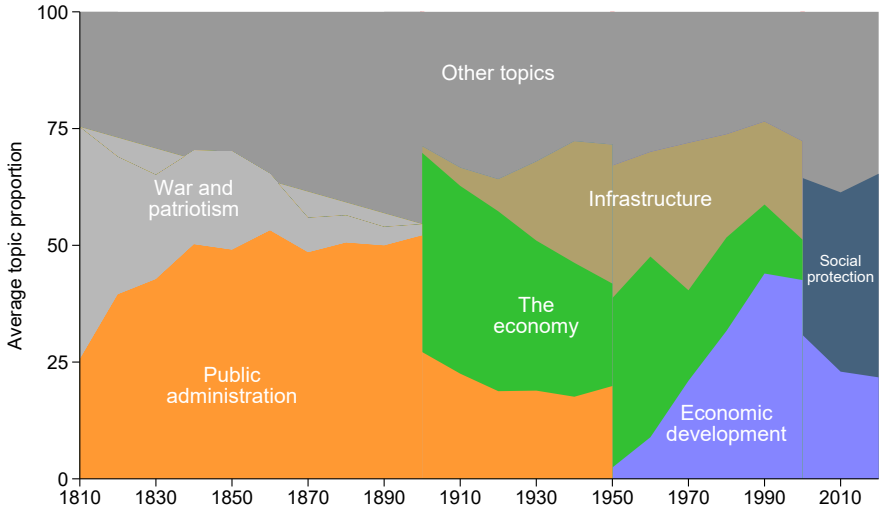

*Notes:* This figure shows the distribution of topics across decades. We estimate the topics and their probability distributions using a Latent Dirichlet Allocation (LDA) algorithm (see Section 3.1 and Appendix D). In Panel A, we set the number of topics equal to 5; in Panel B, equal to 15; in Panel C, equal to 45. We present only topics whose probability exceeds 20% in at least one decade. The rest of the topics are grouped together in the category labeled as "Other topics." Topics are defined by their top occurring keywords (see Table 2 for the top ten keywords that define the topics in the figure). We manually labeled topics based on the top keywords. To construct the figure, we pool speeches at the decade level and

Table C2: Top keywords defining main topics

| | 1819–1900 | | 1901–1950 | | | 1951–2000 | | | 2001–2021 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | War and Patriotism (1) | Public administration (2) | Infrastructure (3) | The economy (4) | Public administration (5) | Economic development (6) | The economy (7) | Infrastructure (8) | Economic development (9) | Social protection (10) |
| 1 | Towns | National | Service | National | Right | Developing | Village | National | National | Family |
| 2 | Homeland | Administration | Plays | Social | Village | Program | Social | Plays | Social | Right |
| 3 | Liberty | Public | National | Production | Life | National | Politics | Construction | Program | Health |
| 4 | Village | Nation | School | Politics | Public | Sector | National | Developing | Developing | Life |
| 5 | Law | Peace | Services | Activities | Executive | Social | Life | Production | Sector | Developing |
| 6 | Honor | Relations | Capital | Plays | Politics | System | Right | Program | Investment | World |
| 7 | Politics | Order | Works | Economy | Nation | Means | Production | Service | Construction | Education |
| 8 | Peace | Right | Construction | Education | Administration | Politics | Economical | Services | Plays | Society |
| 9 | Mens | Plays | Number | Economical | Homeland | Increase | Nation | Bank | Increase | Social |
| 10 | Revolution | Public | Studies | Trouble | Order | Process | Liberty | Plan | Management | Politics |

*Notes*: This table lists the top ten keywords that define the topics in each in Appendix Figure C2. We only list topics whose probability exceeds 20% in at least one decade. Each column header shows the manually-assigned label of each topic. We recover the topics and their probability distributions using a Latent Dirichlet Allocation (LDA) algorithm (see Section 3.1 and Appendix D).

Figure C2: Robustness of topics discovered to conducting LDA by subperiods




*Notes:* This figure shows the distribution of topics calculated separately in four periods (with a similar number of observations per period): 1819–1900, 1901–1950, 1951–2000, and 2001–2021. We estimate the topics and their probability distributions using a Latent Dirichlet Allocation (LDA) algorithm (see Section 3.1 and Appendix D). We fix the number of topics to 25 across all periods (the number of topics that minimizes perplexity for the pooled sample). We present only topics whose probability exceeds 20% in at least one decade. The rest of the topics are grouped together in the category labeled as "Other topics." Topics are defined by their top occurring keywords (see Table C2 for the top ten keywords that define the topics in the figure). We manually labeled topics based on the top keywords. To construct the figure, we pool speeches at the decade level and compute the average topic probabilities in each decade.

# D Latent Dirichlet Allocation Algorithm

## D.1 Description of the LDA Algorithm

Topic models are statistical algorithms that uncover the main topics contained in a collection of documents (Blei, 2012). The most widely used topic model is the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003). The fundamental assumption of the LDA algorithm is that the observed documents were generated through a particular probabilistic generative process.[4] However, the parameters or "recipe" of this generative process are hidden or "latent." This defines the key inferential task of LDA: estimating the "latent" structure—the topics and the topic composition of each document. LDA performs this task by working through the generative process in reverse. That is, it uses the observed words in each document to estimate the parameters of the generative process that are most likely to have generated the observed collection of documents.

The generative process can be described as follows:

1. For each topic, decide what words are likely.
2. For each document:
   (a) Decide what proportions of topics should be in the document, say 20% topic A, 50% topic B, 30% topic C.
   (b) For each word:
      i. Choose a topic. Based on the topic proportions from step 2.a., topic A is more likely to be chosen.
      ii. Given this topic, choose a likely word (generated in step 1).

We can describe this process more formally using the model parameters and corresponding probability distributions. After specifying a number of topics $k$:

1. For each topic $k$, draw a distribution over words $\varphi_k$ according to a Dirichlet distribution $\sim \text{Dir}(\beta)$, where $\beta$ is the parameter of the Dirichlet prior on the per-topic word distribution.[5]

[4] Another important assumption is the "exchangeability" or "bag-of-words" assumption, which means that the order in which the words appear in the document is not important; LDA relies on term frequencies instead.

[5] The beta parameter represents the "prior" belief about the per-topic word distributions. A high beta value means that each topic is likely to be made up of most of the words in the corpus, whereas a low beta means that each topic will have fewer words.

2. For each document $d$:

   (a) Draw a vector of topic proportions $\theta_d$ according to a Dirichlet distribution $\sim$ Dir($\alpha$), where $\alpha$ is the parameter of the Dirichlet prior on the per-document topic distribution.[6]

   (b) For each of the $N$ words $w_n$:

      i. Draw a topic assignment $z_n$ according to a multinomial distribution $\sim$ Multinomial( $\theta$) according to the topic proportion $\theta_d$.

      ii. Choose a word $w_n$ from $\Pr(w_n|z_n, \varphi)$, a multinomial probability conditioned on the topic $z_n$.

The key inferential task of LDA consists in performing this assumed generative process in reverse. That is, using the observed documents and words, LDA works backwards to infer the "latent structure"—the distribution of the parameters $\theta$, $z$, and $\varphi$—that are most likely to have generated the documents in the sample. Where $z$ represents the per-word topic assignments and $\theta$ gives the topic distribution of each document, which indicates the extent to which each document belongs to each topic; $\varphi$ gives the distribution of words in topic $k$, which is used to define the semantic content of each topic. The objective of LDA consists in computing the posterior distribution of these hidden variables given a document and the Dirichlet priors:

$$\Pr(\theta, z, \varphi|w, \alpha, \beta) = \frac{\Pr(\theta, z, \varphi|\alpha, \beta)}{\Pr(w|\alpha, \beta)}.$$

Estimating the maximum likelihood of the model and the distributions of the hidden variables requires marginalizing over the hidden variables to obtain the model's probability for a given corpus $w$ and priors $\beta$ and $\alpha$.

$$\Pr(w|\alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left( \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left( \prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta.$$

These distributions are intractable to compute, requiring the use of other approximate inference algorithms. Although in the first introduction of LDA, Blei et.al. (2003) relied

[6]The alpha parameter represents the prior belief about the per-document topic distributions. A high alpha-value means that each document is likely to contain a mixture of most of the topics, and not any single topic specifically, whereas a low alpha-value means that each document is likely to contain fewer topics.

on a Variational Bayes approximation of the posterior distribution, we use Collapsed Gibbs sampling as our inference technique—a commonly used alternative introduced by Griffith and Steyvers (2004).

### D.2 Collapsed Gibbs Sampling

The Collapsed Gibbs sampling algorithm is a common Markov Chain Monte Carlo (MCMC) algorithm that is used to approximate posterior distributions when these cannot be directly computed. The idea is to iteratively generate posterior samples by looping through each variable to sample from its conditional distribution while retaining the values of all other variables fixed in each iteration (Yildirim, 2012). Essentially, we simulate posterior samples by sweeping through all the posterior conditionals, one random variable at a time. Because we initiate the algorithm with random values, the samples simulated at the early iterations are likely not close to the true posteriors. However, the process eventually "converges" at the point where the distribution of the samples closely approximates the distribution of the true posteriors.

In LDA, the variables we want to approximate are the "latent" variables $\theta$ and $\varphi$. This is achieved by generating a sequence of samples of topic assignments $z$ for each word $w$. As mentioned above, for each iteration, Gibbs Sampling requires retaining the values of all variables except for one fixed (see Griffiths and Steyvers, 2004). Therefore, because words are the only observed variables in LDA, at each iteration, the topic assignment of only one word is updated, while the topic assignments for all other words are assumed to be correct (i.e., remain unchanged). Samples from the posterior distribution $\Pr(z|w)$ are obtained by sampling from:

$$\Pr\left(z_i = K|w, z_{-i}\right) \frac{n^{(j)}_{-i,K} + \beta n^{(di)}_{-i,K} + \alpha}{n^{(.)}_{-i,K} + V\beta n^{(di)}_{(-i,.)} + k\alpha},$$

where $z_{-i}$ is the vector of current topic assignments of all words except the $i$th word $w_i$. The index $j$ indicates that $w_i$ is the $j$th term in the entire vocabulary of words in the corpus ($V$); $n^{(j)}_{-i,K}$ indicates how often the $j$th term of the vocabulary is currently assigned to topic $K$ without the $i$th word. The dot "." indicates that summation over this index is performed; $d_i$ denotes the document in the corpus to which word $w_i$ belongs; $\beta$ and $\alpha$ are the hyperparameters of the prior distribution explained above (Grün and Hornik, 2011).

Intuitively, the algorithm beings by going through each document and randomly assigns

each word in the document to one of the K topics. Because these assignments are random, however, they are poor and must be improved on. To improve these topic assignments, for each word $i$ in document $d$, (for each $w_{d,i}$) and for each topic $k$, two values are computed: 1) Pr(topic $k$ | document $d$), or the proportion of words in document $d$ that are currently assigned to topic $k$, and 2) Pr(word $w$ | topic $k$), or the proportion of assignments to topic $k$ over all documents that come from this word $w$. Then, these two proportions are multiplied to get Pr(topic $t$ | document $d$) $\times$ Pr(word $w$ | topic $t$), which in the context of LDA's generative process, gives the probability that topic $k$ generated word $w$. Finally, word $w$ is reassigned to a new topic based on this probability. To put it simply, for each word, its topic assignment is updated based on two criteria: 1) How prevalent is that word across topics? 2) How prevalent are topics in the document?

As in any Gibbs Sampling algorithm, the above steps are repeated a large number of times. After a large number of iterations, the algorithm converges to a steady state where the topic assignments of each word are close approximations of the true values. At this point, we can finally use these topic assignments to estimate the "latent" variables—the posterior of $\theta$ and $\varphi$—given the observed words $w$ and topic assignments $z$:

$$\hat{\theta}_K^d = \frac{n_K^{(d)} + \alpha}{n^{(d)} + k\alpha} \quad ; \quad \hat{\varphi}_K^j = \frac{n_K^{(j)} + \beta}{n_K^{(.)} + V\beta} \qquad \text{for } j = 1, ..., V \text{ and } d = 1, ...D$$

With $\theta$ and $\varphi$ estimated, the objective of LDA—extracting topic representations of each document—is achieved.

### D.3 Empirical Implementation

As a first step before running LDA, to improve the discovery of topics, we follow the standard practice of cleaning our collection of documents. Specifically, we remove all punctuation and numbers, as well as "stop words"—terms such as articles, conjunctions, and pronouns that are semantically meaningless for defining a topic. Because we are interested in discovering topics that are common across countries, we remove country-specific terms such as "Peruvians," "Peru," or "Lima," which could otherwise bias the topics towards country-specific rather than subject-specific topics.

We rely on Collapsed Gibbs Sampling for the iterative process of topic inference. This approach requires the specification of values for the parameters of the prior distributions—$\beta$ for the per-topic term distributions and $\alpha$ for the per-document topic distributions. Following Griffiths and Steyvers (2004), we select the commonly used values of $\alpha = 50/t$

(where $t$ is the number of topics) and $\beta = 0.1$.

### D.4 Number of Topics to be Discovered

One crucial parameter that must be specified is the number of topics to be discovered. To determine the optimal number of topics we rely on a measure known as *perplexity* that is often used in information theory and natural language processing to evaluate how well a model can predict the data, with lower perplexity indicating a better model (Blei et al., 2003). Formally, for a test set of $M$ documents, perplexity is defined as:

$$\text{Perplexity}\,(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log \Pr(W_d)}{\sum_{d=1}^{M} N_d}\right\},$$

where $W_d$ represents the words in document $d$ and $N_d$ the number of words. The lower the perplexity, the better the model can predict the data.

To minimize the perplexity, we follow a three-step process: (i) Randomly partition our sample into a training set (90% of speeches) and a held-out test set (10% of speeches); (ii) Iteratively implement LDA on the training set using a different number of topics $n \in \{1, \ldots, 30\}$ in each iteration; and (iii) Compute the perplexity of each model, which amounts to evaluating how "perplexed" or surprised each "trained" model is when presented with the previously unseen test set.