Online Appendix:
Data Collection and Measurement

Expanding Executive Authority And Prosecuting Presidents:
An Event Data Study of Developing Democracies

Natalie Ahn
University of California, Berkeley
natalieahn@berkeley.edu

In designing any study, researchers face a series of high-level decisions about how to construct a process that can answer the questions posed. What is the exact phenomenon we wish to study? What sources of information are available to observe it? How can we use those sources to measure the phenomenon of interest? How will we incorporate the measures into analytic models to test for hypothesized relationships, and how do the decisions we've made about data and measurement affect the results we're able to obtain?

In this appendix, I explain the data collection and measurement methodology behind the accompanying paper. This includes the document sources and text processing used to measure leaders' efforts to consolidate power, the event data used to measure the consequences of those actions for leaders after departing office, and the control variables used to capture other explanations from the previous literature on executive power.

A Power-Consolidating Decrees

A.1 Document sources

Researchers studying government action and political institutions often collect data from secondary news media, surveys and polls, or abstract indicators compiled and coded by experts. The latter indices tend to represent static states, often as composite scores that change infrequently and are difficult to relate to similar concepts or decompose into more specific events. Data-driven approaches that make use of more dynamic events have become more common in recent years, with increasingly available digital sources of information. With regard to political actors, event data studies often focus on government action directed outward, such as conflict between states, which are observable in secondary news reporting (Bond et al., 2003; Schrodt, 2006; Raleigh et al., 2010). News reports tend to be informative and straightforward, with the most important details first, and to be useful for monitoring crises and detecting major new developments (Tanev et al., 2008).

However, news media pose challenges to measuring events consistently over time. News reporting is often redundant, so that to accurately count specific events, researchers must deconflict reports about the same incident. News reporting is also influenced by many factors other than whether events occurred, such as the publication's resource constraints, reporters' access to participants, the shifting interests of the target audience, and publishers' business objectives (Kepplinger, 2002; Althaus et al., 2011; Weidmann, 2015; Ortiz et al., 2005). Since

news is designed to sell, it is more useful for studying large-scale attention-grabbing events like new wars or regime changes, rather than everyday policy decisions and interactions.

There are many important research questions, however, that revolve around the day-to-day business of government and the evolution of public offices and authorities. The best or only source of those day-to-day activities may be official government records. Governments are increasingly making their records publicly available in digital archives, including legislation and decrees (Cardie and Wilkerson, 2008). These documents offer new opportunities to study government authority and action in more systematic and detailed ways. Laws and decrees are primary source documents that themselves enact the policy decisions or institutional changes they report. This means that each law or decree constitutes a separate action, which do not need to be deconflicted for redundancy and as a collection are more complete. The language used in government documents is also more formal and consistent.

Researchers still face challenges to collecting and processing large volumes of full text records, especially in developing countries. Despite recent advances, there are often considerable differences in quality, completeness, and usability across sources. Archives may only contain documents' metadata (e.g. titles, dates, and sponsors), or incomplete full text records in the form of scanned images, requiring error-prone conversion to machine-readable text. Laws and decrees are considered public domain in many countries, yet limited resources may impose practical constraints on coverage or ease of access. As with news media, norms are still developing regarding database use for those seeking to mine archives for research purposes, especially in countries outside the United States (Truyens and Eecke, 2014).

As noted in the accompanying paper, this study focuses on executive decrees, which can be attributed directly to the government head. I use decree titles to identify the decisions of interest, since the titles are most widely available across countries and years. Decree titles tend to be declarative statements that identify the gist of the decree's main provision(s), in correct legal terms and with enough detail to distinguish it from similar documents, while fitting on a few lines. These titles tend to have similar linguistic structure across countries,

and can be thought of as combining some of the brevity and directness of news reports with the authority and formality of official records. Several examples are shown in Table A.1.

Table A.1: Example decree titles from project dataset

Date	Title (manually translated)
9/29/2011	Authorize transfer of funding allocation in favor of the Ministry of Defense and Ministry of Interior in the public sector budget for fiscal year 2011
7/6/2003	Approve technical regulation regarding conductors and electrical cables for mass consumption and general use
9/12/2007	Declare Sunday, October 21, 2007 National Census Day for the purpose of the National Censuses of Population (XI) and Housing (VI)

I have collected dates and titles of 73,670 executive decrees issued in the five Andean countries over the past few decades. Most constitutions authorize a single type of executive decree, but I include all relevant subtypes wherever multiple forms of executive decree-making authority are granted. This means that for Peru, I include the three main types of executive decrees (*decretos supremos*, *decretos legislativos*, and *decretos de urgencia*) to maximize coverage of the policy decisions presidents make through decrees, and to make the data comparable to the other countries. Table A.2 summarizes the decrees collected, years covered, and source archives for each country in the dataset.

Table A.2: Decrees collected, years covered, and sources for each country in dataset

Country	Years	Decrees	Source
Peru	1980-2016	30,314	*Sistema Peruano de la Informacin Jurdica, Ministerio de Justicia*, http://spij.minjus.gob.pe
Bolivia	1982-2016	13,927	*Gaceta Oficial de Bolivia,* http://www.gacetaoficialdebolivia.gob.bo
Colombia	1999-2016	16,260	*Presidencia de la Repblica,* http://historico.presidencia.gov.co
Ecuador	2000-2016	9,711	*Registro Oficial de Ecuador,* http://www.registroficial.gob.ec
Venezuela	2009-2016	3,458	*Gaceta Oficial, Tribunal Supremo de Justicia,* http://www.tsj.gob.ve/gaceta-oficial

A.2 Document classification

Studies of executive orders and decrees generally measure their use in terms of the total number issued each year, or the number that meet a certain threshold of salience, which are hand coded from a limited number of government documents and/or secondary sources. Mayer and Price (2002) and Howell (2003) count executive orders as significant based on mentions in news and other political or legal references. Wright (2014) focuses on a small subset of executive decrees in Andean countries that were enacted using emergency powers, and codes whether the decrees cite social unrest and whether they involve the use of force. As discussed in the accompany paper, leaders might have a variety of motivations for issuing decrees. Rather than focus on all decrees, including those that enact temporary emergency measures or deliver one-time programs or services to constituents, I am interested in when leaders use decrees to expand their own institutional power in lasting ways.

For applied event data projects, news and public archives are often too large – and the events or topics of interest too sparse – to be hand coded, requiring some form of automatic classification or information extraction. This is often a multi-step process, involving several key decisions: 1) what categories or labels we seek to encode, 2) what information (features) from the original text to use in assigning labels, and 3) what algorithm or encoding process to use to transform the input features into the output labels. I explored several methodological options for each of these steps, and arrived at an ensemble approach that combines existing tools and resources with limited project-specific rules and procedures, to assign decrees to policy decision categories with reasonable efficiency and accuracy. I define each of the steps below, then evaluate the accuracy and conceptual validity of the resulting data.

Step 1: Coding scheme

The objective for the dependent variable is to identify decrees that are most likely to alter the internal allocation of government authority, and especially to increase the power of the executive in lasting institutionalized ways. For instance, decrees that create new executive

agencies or delegate new powers to those agencies, are more likely to enable subsequent executive actions than decrees that simply distribute one-time goods and services to private constituents. These distinctions involve a bit more than an overall document theme, but do not require identifying every aspect of each decreed event, such as the duration or recipient's location. For each document, I assign one *main action* and one *target entity*, which in combination define different types of policy decisions with different purposes or goals.

Table A.3: Coding Scheme with main action and target entity categories

Category	Subcategory	Example Terms
	Action	
enable	*enable_empower*,	create, delegate, authorize ...
	enable_appoint,	appoint, nominate ...
	enable_finance,	transfer, fund ...
	enable_modify	
regulate	*regulate_restrict*,	require, prohibit, limit ...
	regulate_rules	administrative regulation, bylaws ...
	regulate_modify	
other act	*other_enact*,	execute, distribute, ratify ...
	other_modify	
	Target	
gov (exec)	*gov_executive*,	presidency, cabinet, ministry ...
public (other)	*public_legislature*,	legislature, congress ...
	public_judiciary,	court, tribunal ...
	public_local,	province, municipality ...
	public_personnel,	civil servants, diplomats ...
	public_institutes,	(usually public) schools, hospitals ...
	public_other	
private	*private_business*,	corporation, industry ...
	private_other	youth, workers, voters ...

In the paper, I present hypothesis tests using decrees in just one of the high-level action categories and one of the target entity categories: decrees *enabling* an entity in the *government executive*. In the more in-depth process of exploring, testing and validating potential measures of the dependent variable, which led to that final categorization, I used a

more detailed coding scheme involved two levels: one with a few high-level action and target categories, and another with more fine-grained subcategories of each. Table A.3 shows the full coding scheme, which I use in this section to evaluate the trade-offs among different methodological options, and to validate that certain categories reflect the decisions of primary interest in this study. I also use the full set of lower-level subcategories to characterize leaders' decrees when predicting their future fates, in Section B.3 below.

Step 2: Text preprocessing and feature extraction

The most common features used in text classification are counts of words appearing in each text, regardless of grammar or word order, i.e. "bag-of-words" approaches. Bag-of-words features require very little preprocessing, beyond tokenization (i.e. segmenting raw text into words), which can be done using whitespace for Spanish as well as English text. The main step is to construct a document-term matrix, i.e. a vector of term counts for each document. I use the 1000 most frequent terms in the corpus as the columns in this matrix. Following common practice, I lemmatize words to their root form (i.e. verbs to infinitive, nouns to singular male) and weight the resulting vectors by inverse document frequency, to emphasize terms that are more distinct to specific documents.

I also construct a second set of features that capture more linguistic structure from the text, targeting terms that are likely to represent the actions and target entities of interest. I first apply part-of-speech tagging and dependency parsing, using the Stanford CoreNLP toolkit (Manning et al. 2014), version 3.7.0, with the Spanish language model for the PCFG parser (Klein and Manning 2003; Spanish models by Jon Gauthier). I then use a set of rules to extract each document's main verb, which is usually labeled by the parser as the sentence "root", or attached to an enabling verb like "propose" or "declare", as in "Propose to *create* a new office ...". I use a list of Spanish enabling verbs selected by inspection of the dataset, to distinguish common procedural terms from the active verbs that indicate what the decree actually does. Actions may also be stated in nominal form, as in "Propose the *creation* of a

new office ...". I use a Spanish spell-checking dictionary and morphological rules to convert noun forms of verbs back to their verb infinitive (Rodríguez and Carretero, 1996).

After identifying the main verb in each decree title, I identify that verb's noun objects, to use in labeling the decree's target entity. The targets might be direct objects, as in "Create a *new agency* ...", or indirect objects that appear after a preposition like "to" or "of", as in "Transfer funds *to the ministry* ...". In the latter case, the direct object ("funds") is a general resource, which clarifies what action is being taken (i.e. a financing operation), but not which entity is being funded. The indirect object following "to" (the ministry) is the organizational target of the action. For simplicity, in this preprocessing step, I extract all direct and indirect objects of the identified main verbs, to include as potential target entities in the structured features for a given decree.

Finally, I incorporate lexical information about each noun object using WordNet, a database of over 100,000 word senses with definitions and hierarchical relationships. A Spanish language version of the Multilingual Central Repository (Gonzalez-Agirre et al., 2012) is available through the Open Multilingual WordNet interface in the Python Natural Language Toolkit (NLTK) (Bond and Paik, 2012). WordNet includes many geographic and political unit names, as well as common nouns for government branches, agencies and offices, and other concepts relevant to policy actions. Each word sense in WordNet includes a list of higher-level parent terms (i.e. "hypernyms"), which help identify what type of entity it is. For instance, "congress" is listed under "legislature", which is listed under "assembly", which is a type of the more general term "group". I include the WordNet hypernyms of each noun object in the decree's structured features as well.

Step 3: Encoding models

I have chosen to compare two approaches to assigning a main action and target entity label to each decree. The first is supervised machine learning, a common approach to document classification today, in which a labeled training set is used to learn a function for which

combinations of input features should be assigned to which output label. Popular off-the-shelf tools are available for this type of classification, which simply require enough labeled training data to learn how to accurately replicate human classification. I've chosen to use several common classifiers from the Python scientific computing package *scikit-learn*.

I test Naive Bayes, support vector machines (with a linear kernel), logistic regression, and random forests. These algorithms mainly constitute linear classifiers, with the exception of random forests, which are compilations of decision trees. Some models may work better for certain types of data or features than for others. Since these off-the-shelf classifiers operate at the document level (or on a single vector of features for each labeled text), I run all classifiers twice on the full corpus, once to assign a main action label to each decree and once to assign a target entity label. I repeat this process using the bag-of-words features and the more structured features as inputs, for comparison.

The second approach uses rule-based pattern matching, to deterministically assign certain statements in the text to certain categories, rather than using a learned probabilistic model. The pattern-matching system is in some ways a more traditional approach to automatic event coding. Instead of searching for exact phrases, however, I use the main verbs and noun objects extracted from grammatically parsed text in the last step (i.e. the structured features also used for machine learning classification).

For each label in the project coding scheme, I define one or more rules that contain conditions a decree title must meet to receive that label. Each rule includes a condition for the document's main verb, plus potential conditions for that verb's objects. An object condition specifies a particular dependency relation (e.g. the verb's direct object) and a WordNet hypernym for the noun that should appear in that grammatical position. If a decree title has a main verb matching the permitted verbs for a given rule, and that verb also has a noun phrase with the right hypernym in the right object position, the document is assigned the action category corresponding to that rule. Rules for target entity categories follow the same format, but any verb is allowed, only the noun object conditions matter.

Table A.4: Example rules to match verb-object clauses to action and target labels

category	**subcategory**	**= verb + dependency relation : [hypernym]**
		Action
enable	*enable_finance*	= "financiar" + [any object]
enable	*enable_finance*	= "transferir" + *dobj* : [assets]
		Target
gov	*gov_executive*	= [any verb] + *obj* : [government department]
private	*private_business*	= [any verb] + *obj* : [industry | corporation]

Table A.4 shows example coding rules. For instance, if a decree title contains the verb "*financiar*" (to finance), the decree will be assigned to the action category *enable_finance*. The verb is sufficiently clear on its own that no object condition is needed. If instead a decree title contains the verb "*transferir*" (to transfer), this could refer to many types of transfers, so the decree will only be labeled *enable_finance* if the verb also has a direct object that falls under the hypernym for "assets" in WordNet (such as words for money, funds, resources, etc). For the target entity labels, any main verb will do. If a decree title has a main verb with an object that falls under the WordNet hypernym for "government department" (e.g. cabinet ministry, treasury, etc), the decree will be labeled with the target category for *gov_executive*.

A.3 Evaluating accuracy in relation to human coding

I test all of the document classification models on a random sample of 1300 decrees from the project dataset, which I hand labeled using the project coding scheme. I began by labeling decrees from Peru – the largest sample in the dataset – to develop the project's coding scheme and initial rules, then refined the categories and rules as I incorporated decrees from the other four countries. I ultimately labeled 500 decree titles from Peru plus 200 decree titles from each of the remaining four countries, resulting in a total of 1300.

To ensure that no model has overfit (or simply memorized) the training examples, we need to test for accuracy on a "held-out" test set. For the machine learning classifiers, I use

10-fold cross validation, iteratively reserving a different 10th of the labeled decrees, training on the remaining 9/10ths, testing on the held-out 10th, then averaging the scores from all 10 folds. For rule-based pattern matching, I wrote and refined the rules while hand-coding the first 150 decrees from Peru, then made minor revisions when labeling the first 100 decrees from each of the remaining countries. I froze the coding scheme and rules while hand labeling the remaining decrees. I therefore evaluate the accuracy of the rule-based system on the last 350 labeled decrees from Peru plus the last 100 from each of the other four countries.

In Table A.5, I report scores for precision (correct guesses out of total guesses), recall (correct guesses out of total true labels), and F1-scores (harmonic mean of precision and recall), reported in percentages for readability. The upper half of the table shows scores for assigning the few categories at the higher level of the coding scheme, and the bottom half shows scores for labeling the more fine-grained subcategories.

Table A.5: Accuracy against human coding, comparing classification algorithms

Few high-level categories				*Bag-of-words*			*Structured features*	
Approach	Classifier	P	R	F1 ±95% CI	P	R	F1 ±95% CI	
Supervised	Naive Bayes	46.5	46.5	**46.5** ±1.8	62.6	62.6	**62.6** ±1.8	
machine	Random Forest	80.5	80.5	**80.5** ±1.8	79.4	79.4	**79.4** ±1.0	
learning	Logistic Reg	81.3	81.3	**81.3** ±1.4	79.5	79.5	**79.5** ±2.3	
	SVM	82.8	82.9	**82.9** ±1.5	76.6	76.7	**76.7** ±1.6	
Rule-based	Pattern-matching	-	-	-	77.4	80.9	**79.1** ±1.9	

More low-level categories				*Bag-of-words*			*Structured features*	
Approach	Classifier	P	R	F1 ±95% CI	P	R	F1 ±95% CI	
Supervised	Naive Bayes	25.3	25.3	**25.3** ±1.9	49.0	49.0	**49.0** ±2.2	
machine	Random Forest	70.6	70.6	**70.6** ±1.7	69.2	69.3	**69.3** ±1.5	
learning	Logistic Reg	69.5	69.5	**69.5** ±2.6	70.3	70.4	**70.3** ±2.1	
	SVM	74.9	75.0	**75.0** ±1.9	66.3	66.3	**66.3** ±2.0	
Rule-based	Pattern-matching	-	-	-	70.9	74.9	**72.9** ±2.3	

For both levels of categories, the support vector machine (SVM) model trained on bag-of-words features performs best, achieving F1-scores of 82.9% for the high-level categories and 75% for the more fine-grained subcategories. (The top scores at each level of the coding

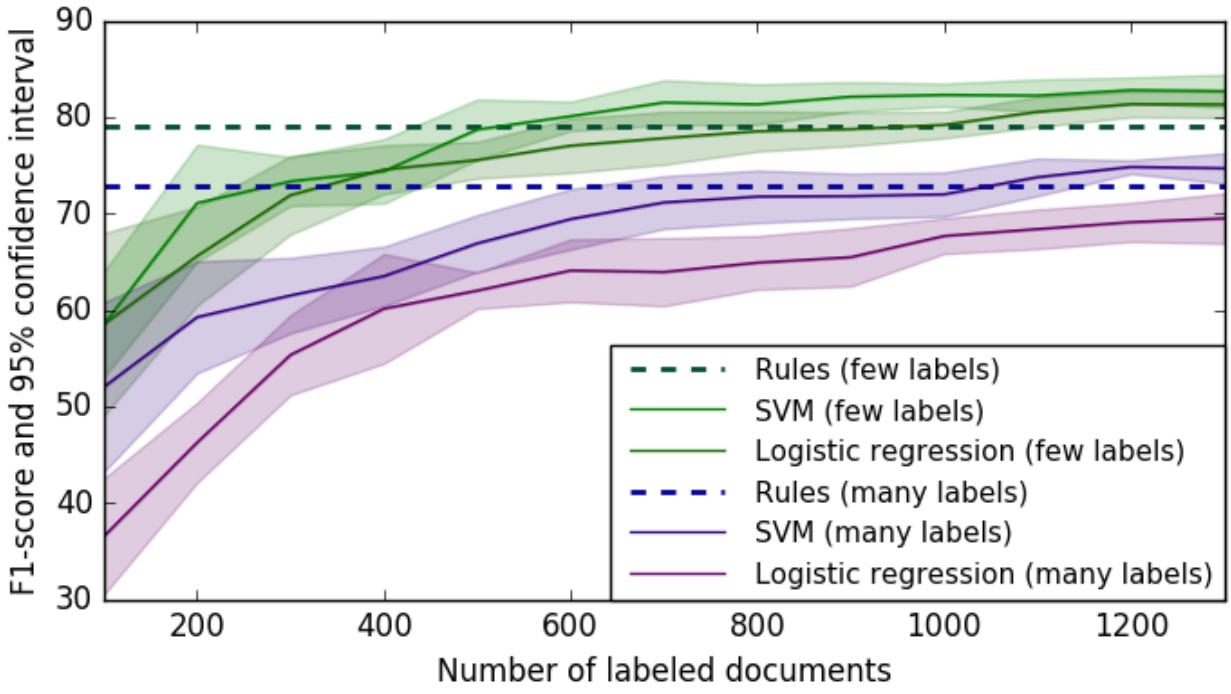
scheme are highlighted in gray.) The random forest and logistic regression models perform similarly on the high-level categories and about 5% lower on the subcategories. Naive Bayes performs much worse, but varies more with the choice of input features.

In particular, Naive Bayes is the only model to perform considerably better when using the more structured features (on the right of the table) than when using simple bag-of-words term frequencies (on the left). Naive Bayes is a fairly simple algorithm that does not account for interactions between features unless explicitly included, so providing specific verb and noun features appearing in certain positions in the text may help to better identify label nuances. The structured features do not appear to help the other classifiers, which seem to have inferred at least as much useful information from the bag-of-words vectors alone.

The rule-based system does not quite match the performance of the SVM model, but it does beat the other machine learning classifiers for labeling the more fine-grained subcategories. The rule-based system achieves an F1-score of 72.9% in the lower half of the table, while the random forest and logistic regression models only achieve scores of about 69-70%. This suggests that even with 1300 hand-labeled documents, there still isn't enough training data for most of the machine learning classifiers to learn how to accurately label all of the coding scheme's more fine-grained subcategories.

Figure A.1 shows how the accuracy of the top two machine learning classifiers increases with the size of the training corpus. For the few high-level categories (the upper lines in the graph), the machine learning classifiers reach the performance of the rule-based system at around 500 labeled documents (for the support vector machine) and 900 documents (for logistic regression). At 1300 labeled documents, both classifiers surpass the rule-based system with statistical significance, using the 95% confidence interval calculated from variation across test folds. However, for the more fine-grained categories (the lower lines in the graph), even with all 1300 labeled documents, the rule-based system still outperforms logistic regression. The support vector machine doesn't surpass the performance of the rule-based system until there are at least 1000 labeled documents for training and testing.

Figure A.1: Accuracy of supervised learning by size of training corpus



Machine learning classifiers need enough labeled examples to distinguish each category included in the coding scheme, so more training documents would be needed to improve the classifiers' scores on the more fine-grained categories. Rule-based information extraction often has lower up-front costs, since it may be easier to explicitly define the most common patterns one is looking for (Chiticariu et al., 2013). However, it may be harder to continue specifying rules for more rare or complex patterns, and easier to hand-label examples of subtler cases, then allow a machine to learn the right combinations of features. With enough training data, as long as the number of training examples is high compared to the number of categories that the machine needs to learn, the machine learning classifiers may be able to achieve higher accuracy than rule-based classifiers on their own.

However, the stronger individual performance of the bag-of-words machine learning classifiers does not mean that the more structured or rule-based options are not useful as well. Computational social scientists have suggested that the best approach to document classifi-

cation is to combine multiple classifiers into an "ensemble" learning process (Grimmer and Stewart, 2013). Table A.6 shows precision, recall, and F1-scores for several combinations of the classifiers tested individually above. For each ensemble, I run all included classifiers on each decree and assign the action and target label with the most classifier votes. Scores that beat the best individual model are shaded, with better scores darker to highlight the gains.

Table A.6: Accuracy against human coding, ensemble classification

Machine learning models included	*Bag-of-words models only*	*Bag-of-words + structured feature models*	*Bag-of-words + structured models + rule-based system*
Few main categories	F1±95%CI	F1±95%CI	F1±95%CI
Best indiv model	*82.9 ±1.5*	-	-
NB, RF, LR, SVM	**81.6** ±1.9	**83.0** ±1.7	**84.2** ±1.5
RF, LR, SVM	**82.7** ±2.0	**83.4** ±1.4	**85.3** ±1.3
More subcategories	F1 ±95%CI	F1 ±95%CI	F1 ±95%CI
Best indiv model	*75.0 ±1.9*	-	-
NB, RF, LR, SVM	**72.4** ±1.9	**72.6** ±1.7	**74.6** ±1.5
RF, LR, SVM	**73.1** ±1.8	**75.7** ±1.3	**76.1** ±2.1

Including all possible classifiers – even Naive Bayes – performs worse than using the better individual classifiers on their own. In the second row of Table A.6, the ensemble with four bag-of-words classifier models only achieves an F1-score of 81.6%, while the best individual classifier achieved 82.9%. Combining just the three high-performing classifiers (in the third row) just about matches the top performing individual classifier. However, the ensembles perform incrementally better from left to right, when including not only the bag-of-words models (on the left), but also models trained on the more structured features (in the center), as well as the rule-based system (on the right). The same is true in the lower half of Table A.6 for the fine-grained subcategories.

This suggests that in an ensemble approach, including multiple models that use different types of information in different ways works better than only using fairly similar classifiers, even if those classifiers performed best individually. Greater variation in the included models

might help take advantage of their different strengths, filling in or compensating for the cases in which the other classifiers are weaker, to achieve the best possible accuracy scores overall.

For the data used in the main analysis in this project, I use the ensemble in the third row and rightmost column of Table A.6, combining the rule-based system with the top three supervised machine learning classifiers, each trained separately on bag-of-words and more structured features, for a total of 7 models incorporated into the ensemble votes. I only use the high-level categories for the main hypothesis tests, since these are classified most accurately, and contain sufficient information for the project's theory (as validated in the next section). This ensemble achieves an F1-score of about 85% on average for all decrees in the hand-labeled test set, which is consistent with many automatic event coding systems used in applied social science research.

A.4 Evaluating downstream contextual validity

While accuracy tests can determine the effectiveness of replacing human coding with an automated process to produce the same output, they can't tell us whether the resulting data capture real concepts of interest for the research objectives. As a final step in the measurement process, I seek to validate that the decree categories capture useful measures of distinct actions leaders take in different contexts, and that one of the categories is a good indicator of leaders' efforts to consolidate power. To do so, I test for expected relationships to other factors that previous studies have associated with the expansion of executive power.

As discussed in the accompanying paper, leaders might increase their use of decrees for multiple reasons, including to simply enact more policies amid urgent demands or gridlock in other decision-making channels. I argue that leaders might take different actions depending on which objective they are prioritizing (whether primarily seeking to enact their policy agenda, or to protect their own interests), and I categorize decrees accordingly. If the resulting categories do a good job of distinguishing the actions leaders take in pursuit of different objectives, we should see the composition of decrees shift among the encoded

categories under conditions that are theoretically associated with the different objectives.

For this analysis, I only use established covariates (i.e. control variables) from previous literature, not the new independent variables about former leaders' fates that I use for the hypothesis tests in the main paper. Since most studies assume that government leaders seek more power to enact their policy agenda, not all prior explanations can be used to differentiate between the multiple motivations identified here. However, there are a few scenarios in which certain measures of established covariates might reveal which types of decrees are associated with one motivation or the other.

The first relates to political opposition and gridlock. When a president's party controls few congressional seats, the president might resort to decrees either to get more policies enacted (i.e. to compensate for bills not passed), or to protect him/herself against potential threats of removal or sanction from the opposition-controlled legislature. In other words, just looking at governing party versus opposition control of congress, the two objectives might be observationally equivalent. The incumbent's concerns about personal threats might vary more, however, depending on how *unified* the opposition is. If opposition parties are highly fragmented, the president might have even more difficulty passing bills, due to the need to negotiate with many smaller parties to secure enough votes. However, highly fragmented parties should pose *less* of a threat of removal or sanction against the incumbent.

If some types of decrees increase as political parties become more fragmented, those decrees might simply represent a shift in the channel being used to enact a policy agenda. If other types of decrees increase only when opposition parties are unified, however, and decrease when parties are fragmented, those decrees might represent efforts to protect the president against opposition threats. For party fractionalization, I use indicators from the Database of Political Institutions (DPI) compiled by the World Bank Development Research Group and the Inter-American Development Bank (Cruz et al 2016). The measures represent the probability that two deputies randomly selected from the legislature will be from different parties, calculated for all parties, and for opposition (non-governing) parties only.

The second scenario relates to near-term pressures when the incumbent has little time left in office, since leaders might shift their strategies as they approach a possible contest for reelection, or a more certain departure from office. If leaders issue certain decrees to accomplish their policy goals, we might see the frequency of those decrees increase whenever the leader has little time left to complete that agenda, whether the leader is seeking reelection or a positive legacy. However, actions that increase executive authority in institutionalized ways should only be observed when incumbents expect to retain the increased power for themselves. If a leader were to strengthen the presidency or executive branch at the end of a term, without a chance of reelection, those changes might only empower a successor.

If we see some decrees drop off at the end of a leader's tenure, then, but only when the leader is not eligible for reelection, we might interpret those decrees as institutional changes that benefit whoever will be in office beyond that point. I code two new variables for months left in office, with and without eligibility for reelection, based on the official term each president was elected to serve and each country's constitutional provisions on presidential term limits. In some cases, presidents were initially barred from reelection, but managed to change the constitution or obtain a judicial ruling allowing them to run again. In those cases, I record remaining months left under the "cannot run" variable up until the legal change, then move remaining months left to the "can run" variable thereafter.

For the analysis in this validation exercise, I aggregate decrees by month into action-target categories, then divide by the total decrees issued that month. By regressing the *fraction* of decrees in each category on the expected covariates, we can observe how leaders change the composition of their decrees under different conditions, beyond changes in the total volume of decrees issued at a given time. Table A.7 shows estimates from bivariate regressions of each decree category (as a fraction of the monthly total) on each established covariate discussed above. I use ordinary least squares with country and year fixed effects, and cluster standard errors by presidential term, since there may be serial correlation between months within the same president's tenure. I do not combine multiple explanatory

factors into the same model here, since they are probably correlated, and testing select combinations of factors can produce unstable results. The goal here is simply to compare pairwise relationships between different subsets of decrees and expected covariates, rather than test a comprehensive theoretical model and interpret the results in absolute terms.

Table A.7: Bivariate regression of decree categories on expected covariates

		all party fractional	*opp party fractional*	*months left in term*	*months left (can run)*	*months left (can't run)*
expected sign for power consolidation		−	−	+	−	+
total decrees		+.262	−.168	−.181 **	−.094	−.246 ***
		(.206)	(.107)	(.085)	(.144)	(.093)
enable	*executive*	−.167 ***	−.138 ***	+.050 *	−.077 *	+.090 ***
		.064)	(.029)	(.027)	(.044)	(.034)
enable	*public*	+.097	+.072	+.034	+.129 ***	−.002
		(.073)	(.057)	(.022)	(.043)	(.026)
enable	*private*	−.059	−.030	−.021	−.016	+.017
		(.059)	(.026)	(.020)	(.014)	(.025)
regulate	*executive*	+.006	+.012 ***	−.003	−.000	−.004
		(.006)	(.004)	(.004)	(.003)	(.004)
regulate	*public*	+.003	−.004	−.007 *	+.007 ***	−.010 *
		(.009)	(.006)	(.004)	(.003)	(.005)
regulate	*private*	−.022	+.007	−.044 ***	−.022	−.046 **
		(.035)	(.028)	(.072)	(.019)	(.023)
other act	*executive*	+.019	+.004	+.014 *	+.033	+.010
		(.015)	(.010)	(.008)	(.021)	(.010)
other act	*public*	−.000	+.008	−.004	−.022	−.008
		(.046)	(.016)	(.016)	(.021)	(.014)
other act	*private*	+.125 *	+.068 *	−.019	−.033	−.048
		(.070)	(.038)	(.026)	(.024)	(.038)

The top row of Table A.7 shows the sign that we would expect to see, for each covariate, in relation to decrees a leader would issue when trying to consolidate power. I've highlighted cells in blue that have coefficients with the expected sign and that are statistically significant, and cells in red that are statistically significant but have the wrong sign (and therefore show the opposite trend). The only row with more than one blue cell is the row for decrees that enable the executive. This is the only category of decrees with a significant negative

relationship to party fractionalization, and especially to opposition party fractionalization. Leaders do not appear to issue more decrees enabling executive offices when we would expect them to be trying to overcome gridlock, but instead when opposition parties are unified and might pose a threat to the president him/herself.

Decrees enabling the executive are also the only ones to diverge in the expected direction as months left in office decline, depending on the incumbent's eligibility for reelection. In most other rows, the coefficients in the last three columns have the same sign. A negative coefficient means that decrees in that category go up (as a fraction of the total) toward the end of a president's term, i.e. when there are fewer months remaining. A positive coefficient means that category is more prevalent earlier in a president's term. Decrees enabling the executive are the only ones that significantly increase toward the end of a leader's term when the incumbent *is* eligible for reelection, but significantly *decrease* toward the end of a term when the incumbent is *not eligible* for reelection.

In terms of the other categories, decrees that enable other public offices (e.g. other branches or local government units), generally move in the opposite directly to decrees that enable the executive. This is intuitive: if those other offices represent rival authorities, empowering them might increase the risk of challenges to the president. I initially considered including decrees that alter constraints on the executive, or that impose regulations on other branches of government, as other forms of power consolidation. However, the *regulating* categories also tend to move in the opposite direction to those enabling the executive. When hand-labeling the training sample, I found it hard to distinguish decrees that impose actual oversight restrictions from decrees that use similar terms for regulation, but simply revise an agency's governing bylaws or enact rules to implement a previously legislated program.

Based on this validation exercise, I only use the decrees that enable executive offices in the main hypothesis tests in the accompanying paper. The analysis in this section provides evidence that the decrees assigned to that category do capture a strong signal that matches the empirical trends we would expect to see when leaders are seeking to consolidate power.

The government executive label was originally a subcategory of all *public sector* actors in the project coding scheme. Based on the analysis in this section, I went back and revised the coding scheme to include not two but three high-level target categories (government executive, all other public sector actors, and private actors). This allows me to use the high-level classification for both decree actions and targets, in the final data to be used for analysis, since the machine learning classifiers are more accurate when encoding only a few high-level labels than when encoding all of the other subcategories as well.

This represents an iterative process in which both the accuracy of the classification algorithms and substantive analysis of relationships to expected covariates informed the final categorization to be used in the main hypothesis tests. It is important to note that the decision about which decree category to use for the final analysis did not involve the new explanatory variables to be introduced in Section B below. This decision was made solely based on control variables derived from pre-existing theories about executive authority. If anything, the use of control variables to select and refine the main dependent variable should favor those pre-existing explanations in any subsequent analysis, setting a high bar for testing my own theory on the same data.

B Leaders' Post-Tenure Fates

B.1 Sources of biographical information

When seeking indicators of government leaders' post-tenure fates, a useful example comes from the Archigos dataset of political leaders (Goemans et al., 2009), mentioned in the accompanying paper. Archigos includes the fates of former leaders in the first year after departing office, labeled as surviving, killed, imprisoned, or exiled. The dataset only contains one fate per leader, rather than a series of potentially conflicting events, and does not include information on other political or legal aspects of those events. Archigos focuses more on violent leadership transitions, and contains more detail on how leaders were removed from

power, including whether government actors, military, or rebel forces participated, with or without foreign support. I focus on attempts to hold leaders accountable for abuses of power in democratic contexts that do not necessarily involve violent conflict, and my theory involves more aspects of how leaders were punished or rewarded over time. My approach is consistent with the general model established by Archigos, however, for studying political leaders' behavior when facing threats to their future interests.

As with the dependent variable, I've chosen to create new measures of the main independent variables in this study, in the form of event data rather than a static state for each leader's ultimate fate. I explored a variety of potential sources, and ultimately chose Wikipedia as the basis for the selection of post-tenure events to include in the project's dataset. English-language news coverage of local events in developing countries is spotty and inconsistent, and local Spanish-language newspapers vary widely in the availability and searchability of digital archives, most only going back a few years. Historic biographies and memoirs are very inconsistent in their coverage of different leaders, and are usually written about sitting or recent presidents, their rise to the presidency and major events in office, rather than what happened to them long after they stepped down.

In contrast, Wikipedia is a widely used free online encyclopedia that is edited by a large and dispersed collective of volunteer contributors, and its accuracy and reliability have been favorably compared to more traditionally edited references (Giles, 2005; Okoli, 2009). The Wikipedia articles on former Latin American presidents contain consolidated summaries of each leader's major life events, including the kinds of information (e.g. dates, locations, and perpetrators) necessary for comparative analysis. While no Wikipedia article includes every single instance of punishment or reward that occurred to a given political leader, I found that the articles shared similar lengths, organizational structures, and types and magnitudes of biographical events mentioned across former presidents.

Working with three undergraduate research assistants with at least some Spanish language skills, I collected data on all post-tenure events mentioned in each leader's English

and/or Spanish Wikipedia entry. We checked the sources cited in Wikipedia to confirm or fill in key details, and searched for other sources (mainly news reports) as necessary, to find exact dates or confirm the nature of the event or other actors involved. The resulting data contains well documented information about major developments in former presidents' post-tenure fates, which were selected for inclusion based on a common, politically neutral reference that provides consistent overviews of public figures' life stories.

B.2 Coding scheme for post-tenure fate events

I have defined the following types of post-tenure events for inclusion in the dataset, summarized in Table B.1 below. For sanctions, I am interested in any attempt to formally punish a former president for his/her acts in office, including through a congressional investigation, judicial inquiry, criminal prosecution, or related actions like stripping a public official's immunity or ordering his/her arrest. For reprieves, I code events that constitute the suspension or abatement of a previous sanction. Reprieves include dismissed charges, acquittal in court, release from detention, or a foreign government's refusal of extradition.

For rewards, I am interested in any formal positions, honors, or other awards that signal former presidents can remain free and enjoy other sources of wealth, status, and influence outside the presidency. As noted in the paper, I only include non-state rewards, since new positions in government or political parties might produce very different incentives for incumbents seeking to use their current office to secure future opportunities. I include former leaders securing leadership roles in business, academia, or other civil society organizations, as well as other formal honors or awards.

Punishing former leaders for abuses of power often involves a lengthy legal process that may play out over many months or even years. Different stages in that process might send different signals to subsequent leaders about the fairness of the process and the certainty of the accused leader's fate. I include a subcategory of sanctions for pretrial arrests and detention, which might represent short-term actions by whoever controls the relevant law

Table B.1: Post-tenure event coding scheme

General fate events	
Sanction	Investigation, criminal charges, prosecution, imprisonment
Reprieve	Dismissed charges, denied extradition, aquittal, pardon
Reward	Major new position or award in private sector or civil society
Sanctions by stages of legal process	
Initial sanctions	Pretrial arrest, pretrial detention
Final sanctions	Formal criminal trial, conviction
Sanctions by predicted probability	
Unexpected sanctions	Low predicted probability based on prior decrees
Expected sanctions	High predicted probability based on prior decrees

enforcement agencies. I also include a subcategory for formal trials and convictions, which usually occur much later, even years after initial arrests are made and charges are filed.

Observable steps in an institutionalized legal process often capture signals about what preceding events have already happened. For instance, reprieves imply that some attempted sanction has already occurred, and is now being reversed. In contrast, convictions usually only occur after investigations have concluded, charges have been filed, and a trial has been held. In this way, convictions may signal a series of *consistent* events, in which multiple actors in the legal system have arrived at a degree of consensus as to the former leader's guilt. Convictions generally entail a higher burden of proof than initial arrest or pretrial detention, and the momentum or consensus that builds throughout the process may convince successors that it will be more difficult to escape justice.

Underlying the hypotheses about different types of sanctions or reprieves is the question of whether former leaders' fates were objective and predictable, on the basis of their prior actions, or arbitrary and subject to political manipulation. While we cannot observe leaders' actual perceptions of these characteristics, it might be possible to construct more direct estimates of the objectivity or predictability of post-tenure fates, in similar ways to the assessments that incumbents might make. Table B.1 includes variables for sanctions broken down by probability, distinguishing expected sanctions from unexpected or surprising ones.

The next section explains the methodology used to calculate those probabilities.

B.3 Predicting the probability of sanction for previous acts

What I am most interested in is how closely former leaders' actual punishments or rewards match expectations about what *should* have happened to them, given how they used their power while in office. We might be able to construct a more direct estimate of this objectivity or predictability, by looking at whether leaders who used their power in similar ways faced similar fates. First, we need to assess what should have happened to each leader, or at least what reasonable observers might have expected to happen, based on what each leader actually did in office. We can treat this as a prediction problem: we see what a former leader did in office and try to predict whether they would be sanctioned, reprieved, or rewarded, given how similar actions by other leaders were punished or rewarded at other times.

Second, we can assess how accurately the classifier predicted each leader's real post-tenure fate. If the predictions are correct for some leaders, we might say that those leaders' fates were unsurprising, which might signal to successors that their own fates will be objective and predictable. If other predictions were *incorrect*, we might say that those leaders' fates were more surprising, which might signal to successors that their own fates may be arbitrary or unpredictable. In other words, if we calculate leaders' predicted fates and then distinguish real events that match the predictions from those that do not, we can test whether incumbents react differently to expected versus unexpected sanctions of former leaders.

For the first step, I use supervised machine learning to calculate the predicted probability of each former leader facing sanction, reprieve, or reward at some point after leaving office, based on the decrees that same leader issued while in power. For training data, I construct one observation for each post-tenure event in the dataset. For input features, I count the number of decrees in each of the project's fine-grained action and target categories that were issued by the leader targeted in the post-tenure fate event, back when that leader was still in office. I also include a feature for the leader's country. For output labels, I assign the general

category of *sanction*, *reprieve*, or *reward* for the given post-tenure event. For simplicity, I include no element of time. The same decree counts are used to predict all post-tenure events that occurred to the same leader. This means that if a leader faced multiple sanctions and rewards, the data will include multiple identical rows of decree counts, but some will be assigned the label for sanction and others the label for reward.

I then use a random forest classifier to calculate the probability of each post-tenure event. To ensure that the resulting scores represent *predicted* probabilities based on each leader's record of decrees, and not each leader's actual (observed) fate, I again use 10-fold cross validation. I iteratively hold out 1/10th of the post-tenure event observations, train the random forest classifier on the remaining 9/10ths, then save the predicted probabilities of sanction, reprieve, and reward from the held-out 10th. For leaders with multiple post-tenure events in the dataset, I average the predicted probabilities of sanction, reprieve, and reward across all of that leader's fate observations. (These predicted probabilities should be the same for a given leader, in expectation, since all of that leader's post-tenure events are predicted using the same decrees as input features.)

Table B.2 shows the average predicted probabilities for sanctions at key points in the legal process. The probabilities are consistent with the interpretation of these different sanctions as discussed in Section B.2. Leaders in the dataset who were detained pretrial, were predicted on average as only 50% likely to be sanctioned, based on their previous decrees. In contrast, former leaders who reached a formal trial and were convicted of crimes were predicted as over 90% likely to face sanctions at some point after departing office. The average predicted probability across all observed sanctions in the dataset was 73.1%. These numbers suggest that former leaders' fates varied in their consistency, and that pretrial detentions may have targeted some leaders whose prior actions were not consistent with sanctions. However, convictions only appear to have happened to former leaders whose prior actions were strongly associated with some form of punishment.

The next step is to incorporate these predicted fates – and how well they match reality –

Table B.2: Predicted probabilities of observed sanctions, by stage in legal process

Type of sanction event	*Mean predicted probability*
Former leader detained pretrial	50.0%
Former leader sanctioned (all)	73.1%
Former leader convicted	92.5%

into the analysis of subsequent leaders' actions. The goal is to analyze whether incumbents react to sanctions against their predecessors differently, depending on how expected those sanctions were. One approach would be to interact observed sanctions with their predicted probabilities, to test for heterogeneous effects over the full range of probabilities. Another approach would be to split the post-tenure sanctions into different variables for events with higher probabilities and lower probabilities, respectively. I try both, constructing two sets of variables as alternative ways of combining observed sanctions with their predicted probabilities, to test for heterogeneous relationships to power-consolidating decrees.

For the first approach, I simply interact the original variable for monthly sanction event counts with the predicted probability of each observed sanction (i.e. the predicted probability that the leader being sanctioned should have been). I first subtract the mean from the predicted probabilities, so that the interaction term contains zeros for sanction events that had an average predicted probability, negative values for lower-probability events, and positive values for higher-probability events. This ensures that lower-probability events are not more closely equated with no sanction at all. Instead, the relationship between the interaction term and power-consolidating decrees should capture how incumbents react to the full range from highly anticipated to highly unexpected sanctions of former leaders.

For the second approach, I split the post-tenure sanctions in the dataset into q quantiles, with roughly even numbers of sanctions (n/q) in each bin, where n = the totally number of sanction events in the dataset and q = the number of quantile bins chosen. To avoid selecting an arbitrary partition of the data, I test a range of values for q. Since there are 49 total sanction events in the dataset, I include quantiles from $q = 3$ to $q = 10$, which means

splitting the data into about 5 to 15 sanction events in each bin. In other words, for a given number of quantiles q, I take the original variable for sanctions by country-month and split it into one variable that only includes the n/q sanctions with the lowest predicted probabilities, another variable with the n/q sanctions that had the next highest probabilities, etc. up to a final variable with the n/q sanctions that had the highest predicted probabilities.

I regress power-consolidating decrees on each quantile of sanctions in turn. Given the different partitions, we then need some way of aggregating and interpreting the results across iterations in which sanction events were split into different numbers of quantile bins. When reporting the results in the paper, I do not show results for all quantiles. Instead, I summarize estimates only for quantiles at certain points along the spectrum of predicted probability. In particular, I expect the strongest relationships to appear at the two extremes, i.e. sanctions of leaders who were strongly predicted to be sanctioned, and sanctions of leaders that were highly unexpected. To show aggregate results at these points on the spectrum, I combine results for the top quantile from each partition into one distribution, then combine results for the median quantiles, and results for the lowest quantiles, respectively.

All regressions still make use of the full 1366 country-months in the dataset, but the explanatory variables have become sparser, with more zeros, since I am only testing leaders' reactions to specific sanctions within certain ranges of predicted probability. The models for high-probability sanctions do not include any of the lower probability events; they simply compare how leaders react to a highly expected sanction of a predecessor, versus how they act in a comparable month when no sanction occurred at all. The distribution of estimates for the highest probability quantiles places greater weight on the most expected sanction in the dataset, since that event is always included in the top quantile. The median distribution places the greatest weight on the median-probability sanction, and the lowest probability distribution places the greatest weight on the very least expected sanction.

C Control Variables

As discussed in the related literature section of the paper, previous theories about increasingly concentrated presidential authority and executive orders generally point to political opposition or gridlock in the legislature, combined with urgent demands for swift government action due to crises or other emergency circumstances, as well as pressure at the end of a president's term. In Section A.4 of this appendix, I used several of these established covariates in the conceptual validation of the project's new decree categories. In this section, I present the full set of control variables included in the paper's main hypothesis tests.

For political opposition and legislative gridlock, I include variables that represent the governing party and main opposition parties' share of seats in the legislature, and variables that represent the degree of fragmentation among parties. These factors might produce different pressures or constraints on executive action, and trigger different concerns about threats to the president. For these variables, I again use the Database of Political Institutions (DPI) from the Inter-American Development Bank (Cruz et al., 2016).

I also include control variables for certain types of crises that presidents might face. I did not use these measures in the validation exercise in Section A.4 above, because their relationship to power consolidation is more ambiguous. While crises might play a role in leaders' efforts to consolidate power, we might see very similar increases in other types of decrees at the same time, that are simply intended to deliver services and alleviate suffering among constituent groups. For that reason, crises are not as useful for validating degree categories, but should be included as alternative explanations in the main hypothesis tests.

As discussed in the paper, executive decrees are often explicitly authorized to enact emergency measures in the face of urgent economic or security crises, but it is difficult to identify domestic macroeconomic conditions or measures of violent conflict that could not have been influenced by the president's own use of power. I instead use global commodity price shocks and natural disasters as more exogenous crises that should influence domestic demands for government action, and to which presidents might react in different ways.

Table C.1: Control variables included in main regression models

Variable	Description	Source
gov seats in legislature	Governing party's share of seats in legislature	Database of Political Institutions (DPI), IADB
opp seats in legislature	Largest opposition party's share of seats in legislature	DPI
party fractionalization	Probability that two legislators will be from different parties	DPI
opposition party fractionalization	Probability that two opposition legislators will be in different parties	DPI
export price chng (1 month)	Top three export commodities' mean price change from last month	Database of Primary Commodity Prices, IMF
export price chng (3 month2)	Top three export commodities' mean price change over quarter	IMF
export price chng (6 month2)	Top three export commodities' mean price change over six months	IMF
export price chng (year-on-year)	Top three export commodities' mean price change over past year	IMF
major earthquake deaths	Count of fatalities from major earthquakes	Global Significant Earthquake Database, NCEI/NOAA
major earthquake injuries	Count of injuries from major earthquakes	NOAA
major earthquake damage	Property damage in millions USD from major earthquakes	NOAA
major earthquake sum score	Sum of codes for earthquakes, deaths, injuries, and damage	NOAA
months left in office	Months left in the incumbent president's mandated term in office	Coded from official presidential terms
eligible for reelection	Whether incumbent is legally eligible for reelection	Coded from each country's constitutional history

For economic shocks, I used data from the International Monetary Fund (IMF) on the global prices of each country's top three primary export commodities. Other types of global economic shocks might have also had considerable influence on executive decrees, such as the financial crisis of 2007-2008. Those types of crises should have hit all five countries in this study at the same time, such that the inclusion of year fixed effects in the main regression models is probably the best means of capturing that variation. For natural disasters, I use data from the Global Significant Earthquake Database maintained by the U.S. National

Oceanic and Atmospheric Administration (NOAA).

Finally, I include the variables coded for remaining months left in the incumbent's term, and for whether the incumbent was eligible for reelection, as introduced in Section A.4 above. Table C.1 shows the full list of control variables included in the main hypothesis tests.

D Statistical Models and Robustness Checks

To combine these variables into hypothesis tests, there are a number of additional decisions to make about model parameters, for which there are many reasonable options that could produce very different results. I have chosen to set up my hypothesis tests using standard approaches to linear regression for cross-national time series data, including some combination of lagged explanatory variables, country and time fixed effects, and robust clustered standard errors (specific factors are discussed below). Filling in this established framework, I take a comprehensive approach to the remaining choices needed to construct specific models for analysis, prioritizing thoroughness and transparency.

Since the number of possible combinations of control variables and other model parameters is very large, I randomly sample permutations, to estimate the distribution of results across as many relevant configurations as feasible. For each independent variable, I construct 1000 regression models, randomly sampling the necessary components. First, I sample control variables to include with the independent variable, selecting the number of control variables from a uniform distribution from zero to all those listed above.

Next, I randomly select the variable on which to cluster standard errors, using either the presidential term or the overall country. I also randomly select whether to include month fixed effects in addition to country and year fixed effects, in case there is any seasonality to the relevant trends. I include at least country and year fixed effects in all models, to account for the fact that the data consist of cross-national time series panels. Naturally, a large portion of the variation across all observations is explained by the country in which the

leader is issuing decrees, since each country has different constitutional provisions for decree-making authority, as well as other differences in executive decision-making structures. Year fixed effects also capture many evolving factors related to the development of democratic institutions, modern mass media, global financial crises, and other trends that are difficult to capture with specific measures, but contribute to substantial changes in the use of decrees over time (Weyland, 2001; Poguntke and Webb, 2005; Green, 2009).

Since my data contain country-month observations, rather than country-year, there might also be some seasonality in the relevant trends. Unlike the legislature, which is usually only in session during certain months of the year, presidents may issue decrees at any point in time. I therefore only include month fixed effects in some models, as an option for robustness. It is also unclear whether we should cluster standard errors at the level of the presidential term, or the overall country. It seems most plausible that decrees may be correlated across months within the same president's tenure, since consecutive months may represent a continuing agenda or policy effort. However, there may be reasons why observations would be correlated across presidents within the same country as well.

Finally, I select a period over which to lag the explanatory variables prior to the month in which the decrees were issued. Researchers often lag explanatory variables by more than one period, if the theorized relationships are expected to play out over longer periods. Since the observations in this study are very short – one month instead of more common country-year observations in comparative political studies – we might expect to observe leaders' reactions unfolding over more than one period. In this project's dataset, the time between any two post-tenure events occurring to former leaders within the same country ranged from less than one month to (in a few outlier cases) several years. About 80% of all post-tenure events in the dataset were succeeded by another sanction, reprieve, or reward of a former leader in the same country within one year.

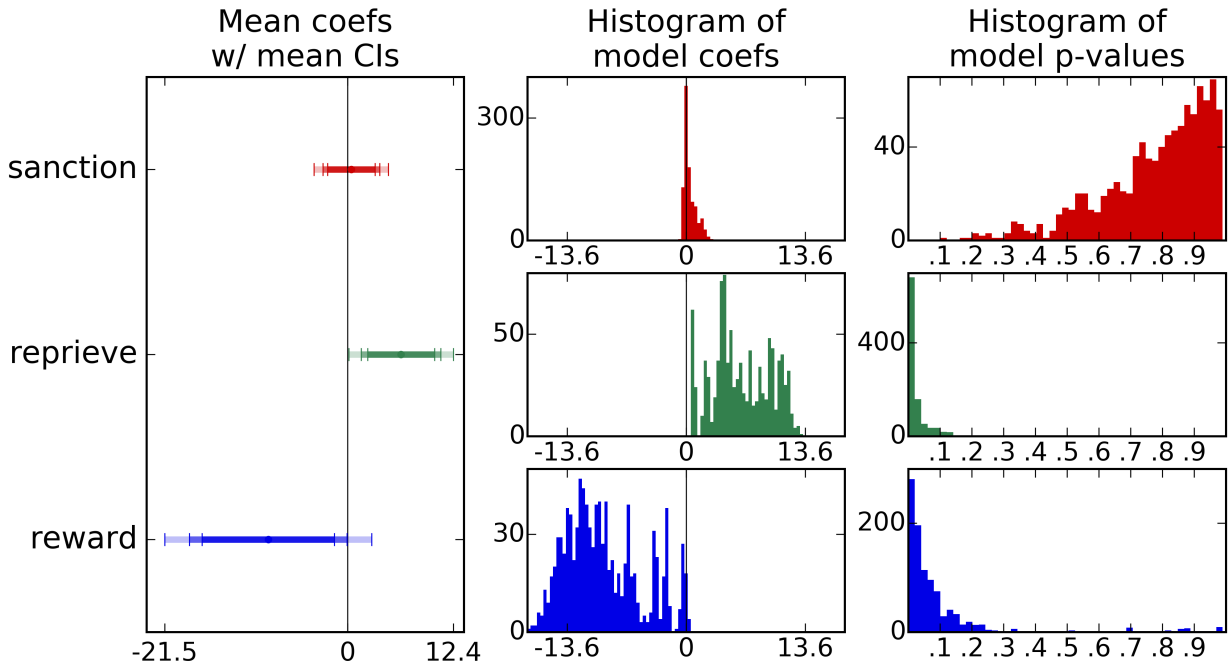
This suggests that in most cases, 12 months should be enough time for incumbents to receive new information about the political environment and potential risks they will face

after departing office. For each regression model, I randomly select a window of n months from a uniform distribution from 1 to 12, making the maximum window consistent with lagged variables used in country-year studies, as well as with the typical periods between post-tenure events observed in this study. I use the average number of post-tenure fate events in each category (e.g. sanctions or rewards of former leaders) over the n months prior to each observation period t, as the explanatory variables for the given model.

For each of the 1000 regression models assembled and run, I record the independent variable's estimated coefficient, standard error and p-value. In the results below, I report average coefficients and confidence intervals across all 1000 regression models. I also plot histograms of the coefficients and p-values, to show the full distribution of results across the different models run. Given the observational nature of the data, testing extensively for robustness to alternative covariates and model specifications can help reduce the possibility of spurious correlation. If any explanatory variable produces consistently significant results across the vast majority of regression models, and if other reasonable explanations have been included as potential control variables, the aggregated results should be highly suggestive of an empirical relationship that is unlikely to have been driven by some other factor.

The figures in the results section of the accompanying paper show the same set of plots and histograms for each independent variable. For baseline reference, Figure D.1 shows the same results as in Figure 1 in the paper, summarizing estimates for the relationship between power-consolidating decrees and general post-tenure sanctions, reprieves, and rewards. When the distribution of coefficients and/or p-values is more spread out, as in the top row for all sanctions, this indicates that the estimates varied more widely with the inclusion or exclusion of different control variables, fixed effects, clustering factors, and/or lag periods. Independent variables with more widely varying estimates might simply be more correlated with the potential control factors, or their relationship to the dependent variable might be more spurious or indirect. When the coefficients and p-values are more tightly grouped, especially when the coefficients are large and all positive or all negative, and when the p-

Figure D.1: Summary of regression estimates for sanctions, reprieves, and rewards (*Separate regression models for each independent variable*)

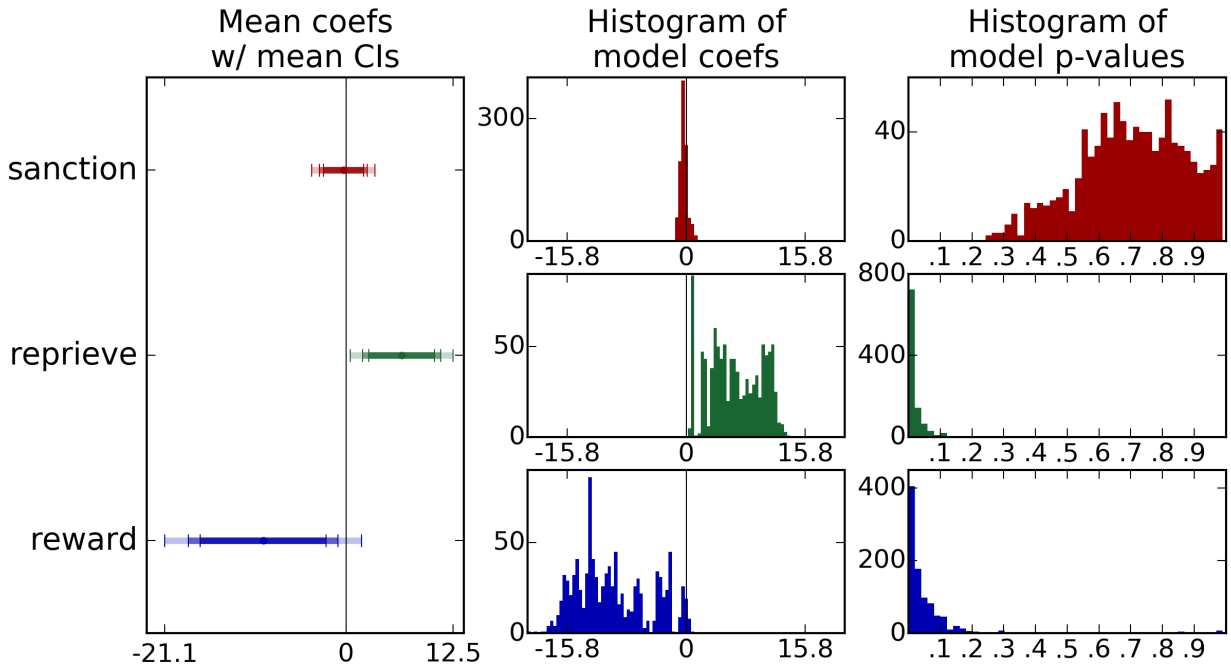


Dependent variable: decrees enabling government executive; 1000 regression models for each independent variable; confidence intervals (left column, inner to outer bars): 90%, 95%, 99%

values are concentrated close to zero, as in the second and third rows, the relationship to the dependent variable appears to be strong and robust to different model specifications.

In the main results presented in the paper, only one independent variable is included in each regression model. For additional robustness, I ran an alternative set of regressions including the main variables for post-tenure sanctions, reprieves, and rewards in the same model. I again sampled 1000 models, including the same potential control variables, fixed effects, clustering factors, and lag periods. Figure D.2 summarizes the estimates for each of these three post-tenure fate variables when included together in the same models. The results in Figures D.1 and D.2 are almost identical; if anything, the p-values have become slightly more significant across the sampled models in Figure D.2, with 91% of p-values below 0.05 for reprieves, and 64% of p-values below 0.05 for non-state rewards. In general, the results are highly consistent; the coefficients are all positive for reprieves and all negative

Figure D.2: Summary of regression estimates for sanctions, reprieves, and rewards (*All three independent variables included in each regression model*)



Dependent variable: decrees enabling government executive; 1000 regression models for each independent variable; confidence intervals (left column, inner to outer bars): 90%, 95%, 99%

for rewards, with very similar magnitude to those in Figure D.1.

The near identical results between Figures D.1 and D.2 lends support to the decision to separately model each independent variable's relationship to power-consolidating decrees (i.e. to choose the approach in D.1, which if anything may be a conservative estimate of the hypothesized relationships.) The fact that the results do not change when we include multiple post-tenure event variables in the same model is most likely due to the sparseness of the post-tenure events, since it is unlikely that a post-tenure sanction and a reward (or even two post-tenure sanctions) will occur in the same country month. The number of sanctions that a particular leader faced at any point after departing office is not independent of his/her rewards, but the exact timing of those sanctions or rewards in relation to each other, does not appear to influence their relationship to power-consolidating decrees. Given the similarity of the separate and combined models, I use models with one independent variable at a time

throughout the paper, for simplicity and interpretability of the results.

The results in Figure D.2 reinforce the conclusions in the paper that former leaders obtaining reprieves from sanction are associated with successors seeking to consolidate power more, while former leaders retiring into non-state roles in society are associated with successors consolidating power less. However, the relationship for sanctions is ambiguous, and whether and how to prosecute former leaders reflects the motivating debate behind this project. In the paper, I provide a series of additional hypothesis tests that break down post-tenure sanctions in more detailed ways. All tests use the same process defined here, of constructing 1000 regression models for each independent variable, randomly sampling control variables, fixed effects, clustering factor for standard errors, and lag period for post-tenure events. The results in the paper therefore capture a high degree of robustness to alternative specifications, while remaining concise and summarizing the prevailing relationship for each hypothesis test.

References

Althaus, Scott L. , Nathaniel Swigger, Svitlana Chernykh, David J. Henry, Sergio C. Wals, and Christopher Tiwald (2011). Assumed transmission in political science: A call for bringing description back in. *Journal of Politics 73*(4), 1065–1080.

Bond, Doug , Joe Bond, Churl Oh, J. Craig Jenkins, and Charles Lewis Taylor (2003). Integrated data for events analysis (idea): An event typology for automated events data development. *Journal of Peace Research 40*(6), 733–745.

Bond, Francis and Kyonghee Paik (2012). A survey of wordnets and their licenses. *Proceedings of the 6th Global WordNet Conference (GWC)*.

Cardie, Claire and John Wilkerson (2008). Text annotation for political science research. *Journal of Information Technology & Politics 5*(1), 1–6.

Chiticariu, Laura , Yunyao Li, and Frederick R. Reiss (2013). Rule-based information extraction is dead! long live rule-based information extraction systems! *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Cruz, Cesi , Philip Keefer, and Carlos Scartascini (2016). *The Database of Political Institutions 2015 (DPI2015)* (IDB-DB-121 ed.). Inter-American Development Bank.

Giles, Jim (2005). Internet encyclopaedias go head to head. *Nature 438*(7070), 900–901.

Goemans, H.E. , Kristian Skrede Gledistch, and Giacomo Chiozza (2009). Introducing archigos: A data set of political leaders. *Journal of Peace Research 46*(2), 269–283.

Gonzalez-Agirre, A. , E. Laparra, and G. Rigau (2012). Multilingual central repository version 3.0. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*.

Green, Jeffrey (2009). *The Eyes of the People: Democracy in an Age of Spectatorship*. Oxford University Press.

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

Howell, William G. (2003). *Power without persuasion: The politics of direct presidential action*. Princeton University Press.

Kepplinger, Hans Mathias (2002). Mediatization of politics: Theory and data. *Journal of Communication 52*(4), 972–986.

Mayer, Kenneth and Kevin Price (2002). Unilateral presidential powers: Significant executive orders, 1949-99. *Presidential Studies Quarterly 32*(2), 367–386.

Okoli, Chitu (2009). A brief review of studies of wikipedia in peer-reviewed journals. *Proceedings of the Third International Conference on Digital Society*.

Ortiz, David G. , Daniel J. Myers, N. Eugene Walls, and Maria-Elena D. Diaz (2005). Where do we stand with newspaper data. *Mobilization: An International Journal 10*(3), 397–419.

Poguntke, Thomas and Paul Webb (2005). *The Presidentialization of Politics in Democratic Societies*. Oxford University Press.

Raleigh, Clionadh , Andrew Linke, Havard Hegre, and Joakim Karlsen (2010). Introducing acled: An armed conflict location and event dataset. *Journal of Peace Research 47*(5), 651–660.

Rodríguez, Santiago and Jesús Carretero (1996). A formal approach to spanish morphology: The coes tools. *Procesamiento del Lenguaje Natural 19*.

Schrodt, Philip A. (2006). Twenty years of the kansas event data system project. *The Political Methodologist, Newsletter of the Political Methodology Section, APSA 14*(1), 2–6.

Tanev, Hristo , Jakub Piskorski, and Martin Atkinson (2008). Real-time news event extraction for global crisis monitoring. *Proceedings of the 13th International Conference on Applications of Natural Language to Information Systems, NLDB*.

Truyens, Maarten and Patrick Van Eecke (2014). Legal aspects of text mining. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

Weidmann, Nils (2015). On the accuracy of media-based conflict event data. *Journal of Conflict Resolution 59*(6), 1129–1149.

Weyland, Kurt (2001). Clarifying a contested concept: Populism in the study of latin american politics. *Comparative Politics 34*(1), 1–22.