Appendix 2 – Technical and Methodological Details

Abstract

The bulk of the work described below can be neatly divided into two sequential phases: scraping and matching. The scraping phase includes all of the steps we go through to gather, prepare, and store the information we want to analyze. Things like extracting information from the web, dealing with any of data's idiosyncrasies that could impact analysis in unwanted ways, and storing the information in a format that makes for easy analysis, are all part of the scraping phase. The matching phase is where actual analysis takes place. In a general sense, this means lining up all of the data we have at the end of the scraping phase, and trying to determine overlap.

Scraping

Federal Register (FR)

We use the Federal Register's API to gather our FR data.[1] Our intended result is a dataset containing all significant rules published in the FR on or after the CRA's enactment into law (3/29/1996). To do this, pass specific conditions to the API's GET::search method, which yields the following URL[2]:

> *http://www.federalregister.gov/api/v1/documents.json?fields%5B%5D=action&fields%5B%5D=page_length&fields%5B%5D=publication_date&fields%5B%5D=regulation_id_number_info&fields%5B%5D=regulation_id_numbers&fields%5B%5D=title&fields%5B%5D=type&per_page=1000&page=1&order=oldest&conditions%5Bpublication_date%5D%5Bgte%5D=1996-03-29&conditions%5Btype%5D%5B%5D=RULE&conditions%5Bsignificant%5D=1*

The API's interactive documentation returns 6683 results total, on seven pages.[3] We construct a loop around the above URL, iteratively appending "1" through "7" to the "page=" portion of the URL string, highlighted above.

Although setting the "Condition" field to "RULE" should—per the Federal Register API's documentation—give us Final Rules only, the "Action" results of our search are not standardized. Of the 6625 results, approximately 54% are labeled "Final rule," with the remainder divided over 803 other unique results. Some of these entries are clearly outside the scope of our inquiry, since they are not themselves final rules that could be reported to Congress. For example, 70 entries have "Final rule; delay of effective date." as their action, indicating an announcement about the delay of the rule's effective date rather than a rule itself. We chose the most frequent action labels that seemed to pertain to finalization of the rule itself,

[1] https://www.federalregister.gov/developers/api/v1/.
[2] Please reach out to nzeppos@brookings.edu for a complete list of our GET::search method conditions.
[3] Note: as more items that satisfy our conditions are published in the FR, the total number of results and pages will increase. This holds true for the GAO, Senate, and House page and result numbers as well.

rather than other procedural aspects of the rulemaking process. This filtering left us with 4350 of the original 6825 results.

Unfortunately, the FR API does not provide data in a ready-made fashion, so we have to clean it up. We first modify it to edit rows that have multiple RIN values. We do this by looping through the dataset and, when we encounter a record of this nature, splitting the RIN values and giving each unique RIN its own row with otherwise equal information.

Next, we handle duplicate RIN values. We search for duplicates and, when we encounter one, preserve only the row that has the highest page length value. We handle duplicate title values similarly—searching for duplicated titles and, when encountered, preserving the record with the highest page number. We eliminate erroneously duplicated rows, and are left with a dataset with unique titles and RINs that is ready for matching against our GAO data.

Government Accountability Office (GAO)
The GAO maintains a database of rules reported in compliance with § 801. Our intended result is a list of all rules reported to the GAO on or after the CRA's enactment into law (3/29/1996). Since the GAO offers this data in a browser-based database instead of an API, we have to use traditional web-scraping methods. We initialize a manual search of all rules reported to the GAO from 3/29/1996 to the present, 50 results per page, sorted from oldest to newest. These conditions preferences give us the following URL:

> *http://www.gao.gov/legal/congressional-review-act/overview?priority=All&agency=All&report=&begin_date=3%2F29%2F1996&end_date=03%2F13%2F2017&end_eff_date=12%2F31%2F2018&end_gao_date=03%2F13%2F2017&rows=50&page_name=fed_rules&path=Legal:Other%20Legal%20Function:Federal%20Rule&searched=1&now_sort=docdate%20asc,issue_date_dt%20asc#fedRulesForm.*

Since there are only 50 results per page, we know we must loop over *n* pages, where *n* = (total number of results / 50). The "About" section of our manual search tells us we have 68,377 total results; *n* = 1367.54, which should be understood as 1367 for our purposes. The GAO link does not contain a "page" field to index over, but it does contain a counter based on your results-per-page preference. The URL for page 2 of our manual search makes this evident:

> *http://www.gao.gov/legal/congressional-review-act/overview?priority=All&agency=All&report=&begin_date=3%2F29%2F1996&end_date=03%2F13%2F2017&end_eff_date=12%2F31%2F2018&end_gao_date=03%2F13%2F2017&now_sort=docdate%20asc,issue_date_dt%20asc&rows=50&page_name=fed_rules&path=Legal:Other%20Legal%20Function:Federal%20Rule&searched=1&o=50#fedRulesForm*

The closest multiple of 50 to 68,377 that does not exceed it is 68,350. Thus, we loop over the by pasting sequential values from 0 to 68350, by 50, appended to the "o=" field of the URL above. On each page, we navigate to each individual rule sub-page, and grab the following information: title, type, description, priority, publication date, RIN.

Senate Executive Communications
Communications from the executive branch to each chamber of Congress are stored in the Congressional Record (CR). Congress.gov has done the work of extracting these executive communications to the Senate. Our intended result is a searchable list of all executive

communication to the Senate from the 104th Congress to present. We want two pieces of information from each executive communication: 1) general identifying information, including Congress of origin and executive communication number, and 2) the text of the abstract; the part of each executive communication that, if applicable, contains a reference to rule reporting.

We first perform an open search for executive communications. Within this search, we set our communication type to "Executive Communications (EC)" and select the 104th – 115th Congress, and ask for 250 results per page. This manual search and subsequent condition modification give us the following URL:

> *https://www.congress.gov/search?searchResultViewType=expanded&q={%22source%22:%22communications%22,%22congress%22:[%22114%22,%22113%22,%22112%22,%22111%22,%22110%22,%22109%22,%22108%22,%22107%22,%22106%22,%22105%22,%22115%22,%22104%22],%22communication-code%22:%22EC%22}&pageSize=250*

This gives us 96,955 results on 387 pages. Adding "&page=" to the end of the URL above gives us an easy way to loop over the pages, resulting in the following URL:

> *https://www.congress.gov/search?searchResultViewType=expanded&q={%22source%22:%22communications%22,%22congress%22:[%22114%22,%22113%22,%22112%22,%22111%22,%22110%22,%22109%22,%22108%22,%22107%22,%22106%22,%22105%22,%22115%22,%22104%22],%22communication-code%22:%22EC%22}&pageSize=250&page=*

Appending page numbers to the "&page=" portion of the URL, highlighted above, provides an easy way to gather our information. We loop over pages 1 through 387, storing relevant identifying information and the abstract for each executive communication we encounter.

To prepare the data for matching, we want to treat each EC's abstract text. We check each abstract for the phrase "rule(s) entitled" and extract the string in quotation marks following this phrase. We then gather identifying information by extracting any text that appears in parentheses in the abstract, as long as it contains at least one numerical character. General identifying information, such as a given executive communication's number and Congress of origin, is inherited from the initial scrape. We store abstracts that don't yield anything from these extractions as "unidentified," and retain the entire abstract to allow multi-step analysis during the matching phase.

The abstract and identifying information data are written separately for spot-checking. Both are then re-read and merged. The resulting merge is our final Senate data set.

House Executive Communications

While Congress.gov stores executive communications to the Senate as separate entries, we must go into the CR itself to find executive communications to the House. To do this, we first gather links to the raw text of every page of the CR that contains executive communications. Executive communications are stored on CR pages that, rather usefully, contain the phrase "executive communications." We initialize a manual search of all pages of the CR this phrase.

Next, we process the text that is found at the pages whose URLs we just gathered. For each body text, we first separate the opening title—a formal introduction denoted by the presence of one of three strings—from the body text. Within the body text, we search for the first instance of

a string calling to the number of the communications, or what can be thought of as the first "entry" in a given CR section. These entries open with a digit, followed by the string "A letter" or "A communication from the President". To make sure we gather each entry, we take note of the number of the first entry, and let the next entry we gather be defined as the previous entry's number plus one. By example: if a given CR section opens with "1400. A letter from the Chairman of the…" we would define this entry as all text beginning at "1400. A letter" until we encounter "1401. A letter" or "1401. A communication from the president . . ."[4] Entry "1401" would be similarly defined by all text up until "1402," and so on. A CR section is defined as ending when the terminal text of an entry is a long series of the "_" character. If no subsequent entry is found and no such string of "_" characters is encountered, we append one to complete the loop.

We define and extract the rule name by first picking a where the rule name likely starts. We do this by checking for the presence of a key word—most commonly "rule" or "order", though we accommodate some exceptions. Plurality is permitted on all key "start" words. We then attempt to guess where the rule name likely ends. We do this similarly, again checking for the presence of some key words—most commonly "received", "pursuant", though we accommodate some exceptions. Once we have the estimated rule title, we search for all text following this rule title that appears in brackets or parentheses. This is our RIN information. Though this way of searching inherits non-RIN identifying information—docket numbers, FRL citations--- our matching process sidesteps this problem.

Matching

FR – GAO matching

We RIN match first. We push parallel vectors of RIN data from the GAO and FR to lower case, strip all punctuation, and remove all whitespace. We look for an exact match from the FR in the GAO and, when encountered, we store the encountered RIN and the location in the GAO vector at which the RIN was encountered.

Next, we look for title matches. A title match, for us, constitutes one of two things: either an exact title match is encountered in the GAO data, or a title match with a maximum Levenshtein distance of 4 is encountered.[5]

This matching process leaves us with 250 rules in our FR dataset seemingly unreported to the GAO. However, we have concerns about the direction of our matching process thus far; specifically, our matching has worked by passing our FR data over our GAO data. To account for this, we look for matches in the other direction: from GAO to FR.[6] Reversing this process does come with some false positive concerns, so we demand that any exact matches found in this iteration have a maximum Levenshtein distance of 20. We find an additional 14 matches title, bringing our total FR rules unreported to the GAO down to 236.

[4] The space in between the digit and "A", in both instances, is optional.
[5] Levenshtein distance is the minimum number of single-character edits between two strings. Our Levenshtein distance is weighted as a simple-edit distance—i.e., all types of single-character (insertion, deletion, substitution) edits are weighted as 1.
[6] In this case, we are only interested for exact matches in our GAO data that may be nested in longer rule titles in our FR data. Distance-matching is non-reversible.

In this and the Senate and House cases, we are looking for a RIN match or a title match to establish proper reporting; either is accepted as sufficient.

FR – Senate matching
We look for FR-Senate matches much like we look for FR-GAO matches.

First, we adjust RIN information from the Senate scrape[7] by pushing to lowercase, stripping punctuation, and removing whitespace. For each RIN in our FR set, search Senate RIN information looking for exact matches. We prepare our Senate rule title data[8] similarly—by pushing to lower case, stripping punctuation, and removing white space—and then, for each rule title in our FR set, we traverse this vector of Senate title information looking for an exact match. When an exact match isn't encountered, we look for the title match with the lowest Levenshtein distance, with a maximum distance of 4.

FR – House matching
First, we adjust RIN information from the House scrape by removing language inherited from previous treatment. [9] Then we parse our House rule data into two pieces: RINs and titles. RINs can be defined as any information nested in parentheses or brackets within our House rule data; titles can be defined as the text remaining once these RINs have been removed from House rule data.

We then proceed to a matching process similar to the GAO and Senate matching. We push all datasets brought into the matching workflow to lowercase, strip punctuation, and remove whitespace. For each RIN in our FR set, search our House data for exact RIN matches; for each title in our FR set, we search for exact matches in our House data. When we don't encounter an exact match, we look for the title match with the lowest Levenshtein distance, with a maximum distance of 4.

Manual correction
Though our treatment is nearly comprehensive, some manual correction is necessary on the back end.

- We return to URLs skipped in our House scraping and manually update our dataset according to the text on these pages.[10]
- Some executive communications—mostly in the 104th and 105th Congress—to the Senate contain language that makes matching tricky, and escapes the matching process

[7] RIN information in this case comes from all parenthetical statements found in all executive communication abstracts. If a given abstract contained multiple parenthetical statements, we combine these and treat them as one string, and then search within it for an exact RIN match.
[8] Title information in this case comes from a string following a specific phrase in all executive communication abstracts. Specifically, rule titles should be understood as: for every executive communication abstract, all strings housed in quotation marks following the phrase "rule entitled", with an optional "s" to append to "rule." Note that this and the RIN parsing are for efficiency and approximate title matching: if matching against these treated data sets yields no matches, we perform an exact match search on the full abstract.
[9] House information, at this point, should be understood as a dataset resulting from parsing all sections of the CR containing the phrase "executive communications".
[10] None of the skipped pages contained executive communications pertaining to rule reports.

we undergo. Most importantly, some executive communications of this variety seem to be reporting rules in large groups. By example, EC1419 in the 105th reads as follows:

> "A communication from the General Counsel of the Department of Transportation, transmitting, pursuant to law, 109 rules including a rule entitled "Establishment of Class E5 Airspace" received on March 13, 1997; to the Committee on Commerce, Science, and Transportation."

Since our program is designed to believe each Senate executive communication relates to one rule, abstracts like the one above are problematic. Unfortunately, other than noting the report of a rule entitled "Establishment of Class E5 Airspace," we run up against the limit of our underlying data. Where executive communications of this nature list RINs, we have tried to correct the Senate portion of our final data. We find the Senate to be the least-reported-to body of the three, and we find a relatively high reporting deficiency rate in these early Congresses. Executive communications of this nature, rather than actual reporting deficiencies, could drive this finding.

Supervised n-gram matching

Initially, we considered the above process to be relatively comprehensive. We were left with over 500 rules that seemed unreported. But, manual checking of our results led us to conclude that, although accurate in identifying matches, our matching process was insufficient. This became particularly apparent when manually investigating the report rules promulgated by the EPA, though it seems to have affected rules reported by other agencies as well. By example:

Our initial matching analysis concluded that the rule "National Emission Standards for Hazardous Air Pollutants; Final Standards for Hazardous Air Pollutant Emissions From the Printing and Publishing Industry" was not reported to the Senate. We didn't find Its RIN, "2060-AD95", by exact matching, and we didn't find title by exact or approximate matching. However, upon manually searching for the rule in a parsimonious fashion, we find the following rule title: and "National Emission Standards for Hazardous Air Pollutant Emissions: Group I Polymers and Resins; Marine Tank Vessel Loading Operations; Pharmaceuticals Production; and *The Printing and Publishing Industry*" (italics added). Given the vagueness with which many executive communications are written, the recognition that language is often quite messy, and our concern with false negatives, we judged it prudent to count titles *like* this as title reports.

In order to define likeness, we employed a flexible search method that used each rule title's most unique words to iteratively search through our data. We defined word uniqueness in two ways, by length and by number of appearances in our dataset. We demanded each rule provide a minimum of 3 unique words, and allowed up to 7 to be included. The number of words used was determined by the length of the rule name, and the number of sufficiently unique words in the rule name. We then searched through each of our underlying data sets—GAO, Senate, and House—by these words, iteratively and by decreasing uniqueness, until we arrived at one or more possible matches. For each of these matches, we performed three additional match steps, in sequence. 1) The non-unique words from the original title name were folded back in, and an exact search for each was performed on the abstract. If a majority of these words were found, the title was counted as a match. 2) If a majority of these words were not found, a date comparison was made. If the FR publication date of the rule in question was within one year of the date of the communication, the title was counted as a match. 3) If neither of the previous steps was satisfied, the identifying information of the first match was stored, and subsequently

hand-inspected to determine whether or not it should be included. This process brought our final tally of unreported rules to 348.

FAQ

Are you reading and counting actual reports?
This is an important point to make upfront: no. The CRA mandates that reports of rules to the GAO and both chambers of Congress.[11] We are not reading or counting reports of this nature. We are programmatically reading and counting publicly available records of short communications from the executive branch to Congress concerning a specific rule or rules, and programmatically reading and counting a database of rules reported to the GAO. This means we are proxy-measuring rule reporting pursuant to the CRA. The CRA demands that reports of rules include specific information, such as "a complete copy of the cost-benefit analysis."[12] The extent to which a report complies with content-specific CRA requirements is beyond the scope of our investigation.

Do you worry about underlying data completeness?
Yes. And we have tried to make this point forcefully. It is conceivable that an executive communication indicating the report of a rule simply didn't make it into the CR, or that the GAO didn't manage to input every single reported rule it received into its database. Importantly, the CRA does not require evidence of a reported rule be documented in either of these ways. This means that unreported rules, as we have measured them, are more accurately characterized as "apparently unreported rules." Our results sketch out a ceiling of the number of rules eligible for repeal under this broad interpretation of the CRA, and should not be understood as a definitive list of all rules eligible for repeal under this broad interpretation.

What are your main concerns regarding false negatives?
False negatives continue to be our primary concern in this dataset. We have made programmatic decisions at multiple steps in our workflow that could have excluded matches. These decisions include but are not limited to: demanding a rule title be written in the CR in a particular way to be included in our exact and simple-edit matching process; demanding specific language be used in executive communications to the Senate to be included in our exact and simple-edit matching process; capping the second tier of our title matching process at a simple-edit distance of 4; requiring exact matching of RINs. None of these decisions were made on a legal basis: a typo or difference in excess of 4 simple-character edits in a rule title would not legally constitute false reporting. Additionally, our concerns regarding underlying data completeness directly relate to the question of false negatives. Put simply, if a report of a rule was never recorded in any of our data sources, it wouldn't have an opportunity to undergo matching analysis. The exclusion of a rule report from any of our underlying data sources does not mean that rule was unreported, and as such it could not form the sole legal basis for undertaking a joint resolution of disapproval.

None of these concerns seems to have amounted to systemic exclusion from any specific Congress or agency: excluding early years, for which we find rule compliance seems evenly distributed across time, and (other than the State Department) no single agency has a strikingly high reporting deficiency rate.

[11] 5 U.S.C. § 801(a)(1)(A).
[12] 5 U.S.C. § 801(a)(1)(B).

Do you have any concerns about false positives and using Levenshtein distance?
We consider our requirements for matching—from the series of data processing decisions to the burden for qualifying as a match—as imposing a relatively high threshold for being included in our final data. As such, false negatives were our primary concern. Though we rate the relative likelihood and impact of false positives as low, they are still a possible source of inaccuracy, and thus a concern. One particular false positive possibility deserves special attention: erroneous year values. A Levenshtein distance of 4 accommodates complete replacement of year values. That is to say—any year value, by virtue of being 4 characters long, is at most 4 simple character edits away of any other year value. By example:

> "Light Truck Average Fuel Economy Standard, Model Year 1998," RIN: 2127-AF16, appears in our FR dataset. We find GAO RIN and title, House RIN and title, and Senate title matches. Our GAO and House title matches are exact matches, but our Senate title match has a Levenshtein distance of 4. The underlying Senate title is "Light Truck Average Fuel Economy Standard, Model Year 2002."

Because this title is our Senate data's lowest distance-match, and its edit distance does not exceed our maximum allowed distance of 4, it is counted as a Senate title match. In this case, a simple edit distance of 4 can be in various sequences, but here is one such example: 1998 → 2998 → 2098 → 2008 → 2002.

Often, allowing for matching at a distance of 4 results in an accurate match that would have been otherwise excluded. By example: "Distance Learning and Telemedicine Grant Program," RIN: 0572-AB22 appears in our FR dataset. We find GAO RIN, House RIN and title, and Senate title matches. Our House title is an exact match, but our Senate title match has a Levenshtein distance of 4. The underlying Senate title is "Distance Learning and Telemedicine Loan Grant Program."

Remember, all of our titles are pushed to lowercase and stripped of punctuation and whitespace, so the addition of the word "loan" would constitute 4 simple-character edits by simple insertion. Since this seems like a match we'd like to be counting, a maximum distance of 4 seems to have paid off in this instance. Importantly and as mentioned earlier, false negatives were our primary concern. Given we had strongly gated entry into the matching process, we deemed it worthwhile to somewhat relax the threshold for matching.

In general, most of our matches are exact or nearly exact matches: approximately 96% of our FR-Senate title matches, 94% of our FR-House title matches, and 97% of our FR-GAO title matches had a Levenshtein distance less than or equal to 1.[13] These statistics provide some reassurance that our title matching is sufficiently accurate, and indicate that our reporting numbers are by-and-large built on robust matching.

What are your concern regarding your n-gram matching process?
Unlike concerns with our exact and approximate matching processes, our concerns with the supervised n-gram matching is a mix of worry about false positives *and* false negatives. More specifically, though our concerns regarding false negatives persist, we believe this final matching step represents the least demanding portion of our matching phase. Some of the results yield matches that are distant—a senate communication three years after a publication

[13] These percentages excluded title matching by n-gram, and count nested exact matches as having a Levenshtein distance of 0.

date. It's important to note that this doesn't happen *only* with n-gram matching; often, RIN matches yield results that are reported long after the original rule's publication date. We have attempted to supervise what we consider the loosest-but-nonetheless-apparent evidence of a match: by hand comparing the closest n-gram match that doesn't satisfy any of the programmatic steps, we hope to have successfully captured the false negatives that escaped our exact and approximate matching steps *while also* limiting false positives.

We are happy to discuss our methods or share our code with anyone interested. Please reach out to nzeppos@brookings.edu with any inquiries.