Appendix A: Data Description

for

"Diversity Effects or Dissent Aversion? Identification and Estimation in Judicial Panel Voting"

Charles M. Cameron
Center for the Study of Democratic Politics, Princeton University

Lealand Morin
College of Business, University of Central Florida

Harry J. Paarsch
College of Business, University of Central Florida

June 3, 2021

Appendix A: Data Description

The Westlaw Database

The primary raw source of data is Westlaw databes, which includes a collection of documents describing cases heard in the U.S. Courts of Appeals. We collect information for all cases that contain the term "sexual harassment" that are categorized under "Labor and Employment" over the twenty-year period spanning the years 2000 through 2019. The court documents are available in several formats: Microsoft Word 97/2003 `doc` format, rich text `rtf` format, and `pdf` format. We download the files in `doc` format and translate them to `txt` format, using the Python module `win32com`.

Our data-collection strategy is built upon on the systematic structure of these `txt` files. There are a few variants of documentation for different case types, and some changes in format over the years, but the information is mainly organized with information listed in the same order and written with stable patterns in the text.

We wrote a Python module called `legalbeagle`, which tracks down information from the text in court documents, and fetches it to be stored as a row in a data frame. The data in each line of text are categorized into one of several fields using functions of the form `is_[field name](line)`, which identifies whether the line of text matches the characteristic of one of the fields of interest. These functions are used either within a function of the form `get_[field name](file, last_line)` or to trigger a call to such a function. These functions, in turn, either read from the file or continue from the last line read by the previous function to parse a field from the court document.

Some of the fields are described in a number of lines that varies across court cases, so the function reads until another variable type is recognized, signaling the conclusion of the data collection for the previous variable. For this reason, other types of fields are collected to ensure the reliability of the collection of the contents of subsequent fields, even if some of these fields are not directly used in estimation.

An example will clarify the data collection process. The most general format if a file, once converted into `txt` format, begins as follows:

```
KeyCite Yellow Flag - Negative Treatment
  Distinguished by Wells v. Hi Country Auto Group, D.N.M., November 13, 2013
656 F.3d 1277
United States Court of Appeals,
Tenth Circuit.
Christie HELM, Plaintiff-Appellant,
v.
State of KANSAS, Defendant-Appellee.
No. 10-3092.
|
Sept. 7, 2011.
Synopsis
Background: Administrative assistant brought action against state, alleging sexual harassment c

Holdings: The Court of Appeals, Ebel, Circuit Judge, held that:
```

```
[1] judge was not alter ego of the state;

[2] judge's sexual harassment of assistant did not culminate in assistant's termination;

[3] state exercised reasonable care to prevent sexual harassment;

[4] state acted reasonably to correct harassing behavior in response to assistant's complaint;

[5] assistant unreasonably failed to take advantage of preventive or corrective opportunities p

Affirmed.

Procedural Posture(s): On Appeal; Motion for Summary Judgment.


West Headnotes (13)


[1]

Civil RightsPractices prohibited or required in general;  elements
CivilRightsHostile environment;  severity, pervasiveness, and frequency


Actionable sexual harassment under Title VII includes not only economic or tangible discrimina
3 Cases that cite this headnote

...
```

The first lines are blank, corresponding to the upper margin in the original Word document in `doc` format. Although not interesting in itself, this first non-field illustrates a source of irregularity in the data: consecutive fields are occasionally separated by an unknown number of blank spaces. Furthermore, some fields span multiple lines and some span a variable number of lines.

For example, the next line begins with `KeyCite Yellow Flag` and is an addition from Westlaw, which indicates that this case was referenced in a later case. The next line `656 F.3d 1277` is an identifier that we, perhaps incorrectly, call a `case_code`. It is a sequence of three strings separated by spaces; the first and last are sequences of digits and the middle string is an alphanumeric code. The next line is `United States Court of Appeals,` (comma optional) which is followed by the circuit number, such as `Tenth Circuit`. In most files, these fields are listed in a fixed sequence in the first five nonempty lines.

The next several lines list the parties involved in the case. In the simplest form, as in the example above, the plaintiff-appellant is named, complete with this labeling, followed by a line containing only `v.` On the next line, the name of the defendant-appellee is listed and labeled similarly. Some court records list multiple parties before or after the `v.` and some court records list multiple parties separated by multiple `v.`'s. In some cases with multiple parties, the multiple

parties are listed on multiple lines but many others are listed within a single line. Although this may be useful information for each case later on, we collected the list of names of parties but did not separate them, as we do not yet have a need for this data.

Following the list of parties—in all files—is the case number, which is a unique identifier that can be used as a key to join with data from other databases. The case number is commonly written as above, in the form `No.  YY-1234`. There exists some variety with which this information is listed. For some files, the case number is written as either `Docket No.  YY-1234` or `No.  YYYY-1234`, or, without the label `No.`, as simply `YY-1234` or even `YYYY-1234`. In some files, the case number has the suffix `-cv` appended, as in `No.  YY-1234-cv`. In other files, multiple case numbers are listed, as in `Nos.  YY-1234, YY-4567`, and a few files have many case numbers listed. Later versions of the `legalbeagle` module will include functions for parsing the case number in the form `YY-1234` to join with information in other databases.

After the case number is often a single line with a pipe symbol, `|`. The next field is a date. In the simplest form, as above, only a single date is listed. If the sequence of events in the case took place over multiple days, these events and dates will also be listed on the following lines, often separated by blank lines or another `|`. For example, one case lists the following:

```
|
Argued: Jan. 13, 2011.
|
Decided and Filed: June 28, 2011.
|
```

Since we have no immediate need for this information, the dates are collected and stored only to continue the flow of the program through the file.

The next line describes the case in sentences. The first line of this description contains only the word `Synopsis`. On the next populated line, the case is described in sentences following the header `Background:`, although, in the files from cases heard before 2004, the text of the background is written without the header `Background:`. Still, it is easy to collect this information because it usually appears in one line of text, even though it often spans multiple lines with word-wrapping. We have no plans to parse any information from this field because it would be more difficult, due to the unstructured nature of the field, however, we might find a need for this information later. Note that the `Synopsis` and/or `Background:` lines do not appear in every case: some cases are judged *per curiam* and the case file is abbreviated considerably. For this reason, and perhaps several others, the background information is not verified as being recorded for a material fraction of the cases. Perhaps as much as ten percent of cases do not have the background recorded, although accurately measuring this fraction is problematic because of the unstructured nature of the field, when the header `Background:` is omitted.

After a space, is a sequence of double-spaced points describing the holdings. The holdings are preceded by a single line in the form above, as in

```
Holdings: The Court of Appeals, Ebel, Circuit Judge, held that:

[1] ...
```

The next several lines are a sequence of statements, enumerated in square brackets, as in `[1]` above. After the last numbered point, the next line is blank and is followed by a statement of the outcome

of the case[1]. In many cases, this is simply the word `Affirmed.` In others, the outcome of the case takes on a hybrid form, such as

```
Affirmed in part and reversed in part.
```

A complete listing of the outcomes of the case is listed in Table 1. We should think about how we

Frequency	Outcome
9,999	Examples
1,000	`Affirmed.`
250	`Affirmed in part and reversed in part.`
1,000	Other

Tab. 1: Outcomes of Cases in U.S. Courts of Appeals

use this field to determine which cases to include in our analysis and how to categorize them.

The holdings and outcome is followed by a statement of procedural posture. In the simplest cases, it may take on the form

```
Procedural Posture(s): On Appeal; Motion for Summary Judgment.
```

In other cases, there may be several items listed in this field, such as

```
Procedural Posture(s): On Appeal; Motion for Summary Judgment;
Motion for Judgment as a Matter of Law (JMOL)/Directed Verdict.
```

These are listed in a single line of text but the second statement is shown above on a separate line to show the added item. I don't understand these terms enough to know what to do with them but it seems as though they are structured in a such way that it will be easy to parse into separate categories: the items are separated by semicolons and the items take on only so many values.

The next section is often a lengthy listing of quotations from legal documents. It is a numbered list of notes under the heading `West Headnotes (X)`, which indicates the enumerated list of notes from `[1]` to `[X]`. The current version of the `legalbeagle` module skips this section.

A few pages later, the next section begins with the header `Attorneys and Law Firms`. A typical example takes on the following form:

```
Attorneys and Law Firms
*505 ARGUED: Justin S. Gilbert, Gilbert, Russell, McWherter PLC, Jackson, Tennessee, for Appel
Before: MERRITT, ROGERS, and WHITE, Circuit Judges.

MERRITT, J., delivered the opinion of the court. ROGERS (pp. 513-14), and WHITE (pp. 514-20), .
```

The first line contains a list of attorneys and law firms representing the plaintiff-appellant. The second line is usually contains a list of attorneys and law firms representing the defendant-appellee. This passage often spans three lines but sometimes the attorneys are listed in a single line. In

[1] I suspect the technical term for this outcome is "the verdict." Nevertheless, I think we should go through the exercise of identifying the proper terminology for all the fields in the court documents, including the term "court documents."

any case, the last line in this section is especially important for our research question, since it lists the names of the judges in the judicial panel. This line usually takes the form shown above, as in `Before: MERRITT, ROGERS, and WHITE, Circuit Judges.` It might, however, list two judges as, for example, `Circuit Judges` and a third judge with another title. The structure of this sentence is standardized enough that it should not be too difficult to separate the names of the judges. The judges' names are often—but not always—stated in upper case letters. The names are sometimes listed with first and middle initials and sometimes with first names and middle initials. It is a reasonable possibility that more that the last names of judges will not be unique. For example, we will have to distinguish between judge

It is possible, but I have not verified this claim, that judges who share the same surname are listed with initials or first names. The next step is to compile a master list of the unique names of judges on these judicial panels and attempt to match it to the information in the database of judges. We have not yet collected information for judges.

A related set of information is the list of opinions. It is typically labeled as the opinion of one of the judges and may be followed by the opinions of some of the other two judges, particularly in the case of a dissenting opinion. The opinions are often written with excerpts from oral arguments or testimony and also excerpts from other legal documents. For now, this information is skipped, since it takes on an irregular format, however, it is worth investigating in order to characterize the outcomes of cases with partial verdicts, such as `Affirmed in part and reversed in part.`. For our application, it matters whether the "reversed in part" part is a result of a disagreement between the judges or a unanimous decision to reverse part of the verdict in the trial case that was appealed.

For instance, the case in file number `088` heard in 2011 has the outcome:

```
Holdings: The Court of Appeals, Merritt, Circuit Judge, held that:

[1] employee filed "charge" with Equal Employment Opportunity Commission (EEOC), ...

[2] supervisor's derogatory statements to employee were based on race, ...

[3] other adverse treatment that employee suffered was not race-based; and

[4] supervisor's statements were not sufficiently severe or pervasive standing alone ...

Affirmed in part and reversed in part.

Rogers, Circuit Judge, filed an opinion concurring in part and dissenting in part.

Helene N. White, Circuit Judge, filed an opinion concurring in part and dissenting in part.
```

We will have to carefully consider how we categorize this sort of case. An important next step is to tabulate the frequency of each outcome.

In some documents, the judges' opinions constitute the bulk of the court document. In others, the opinions are briefly stated, often in a single sentence, stating little more than the verdict.