Analyzing Federal Prosecutorial Misconduct

Ben Gaudiosi, Hao Chen, Justin Ingwersen

Introduction

Reporters have heard rumors from their colleagues and the opinions of judicial experts of an epidemic of prosecutorial misconduct, specifically with respect to drug enforcement and immigration. However, the data to back this up is obfuscated behind a complex government dump of files that is difficult for reporters to analyze. This dataset is called the Legal Information Office Network System, or LIONS.

In this project, we have developed a tool that can be run to organize the posted text files into a database that is easy to query and accessible to any reporter. In addition, we performed analysis on this data using various classification techniques to determine whether or not a future case has a good chance of being thrown out due to prosecutorial misconduct. To accomplish this, we created a tool in Python that reads these monthly dumps and parses the data in order for us to analyze it.

Datasets

Our primary dataset we used was the LIONS database, which can be found at https://www.justice.gov/usao/resources/foia-library/national-caseload-data. The dataset is distributed from the website in the form text files, where each file represents a table in the database. The text files are distributed with the schema for each table, likely from a MySQL database. We used these schemas as well the LIONS appendix to recreate the LIONS database in PostgreSQL so that we could query it on our own, determine various codes used by the DOJ, and finally hand off to our partner to continue researching the data as our project concludes.

Using these schemas, we found the important information we needed from the database. There is a main table called ‘gs_case’ which contains most of the main elements of a case, such as U.S. Federal District, type of case, important dates, and other details of the case. We found several other tables in the dataset containing related data, which we then combined to create a feature set for our data. With this feature set, we could now attempt to predict the outcomes of cases in the U.S. criminal justice system in order discover cases of prosecutorial misconduct.

Algorithms & Analysis

Using the above feature set, we created both a Decision Tree classifier as well as a Logistic Regression model to determine the outcome of a case.

First, we had to transform our data in order to be properly classified and analyzed. The files were often several gigabytes in size. Therefore, in order to successfully load the large amounts of data onto our systems, we took a sample of the feature set - approximately ⅓ of the total data - and ran our classifying techniques and algorithms on this set. Once we had a dataframe for each of our 'gs_case', 'gs_court_hist', 'gs_court_judge', and 'gs_judge' data files, we merged them together on the basis of the case ID as well as the district the case was in. We needed to include the district in the merge because we discovered that multiple districts in the US use different case ID systems and thus there are duplicate case IDs when looking at the US as a whole.

The next step was formulating a training dataset where we knew or there was a high probability that a case experienced prosecutorial misconduct. To do so, we looked through our data for a field called "Disposition Reason" which details the final settlement of a case. Working with our partner, we handpicked a list of disposition reason codes from the tables that were the most clear indicators that misconduct was involved. For example, one code we were looking out for was 'LECI', which stood for Lack of Evidence of Criminal Intent. Another was 'WKEV', which stood for Weak or Insufficient Admissible Evidence. According to our partner Brooke, these codes often appeared in cases where there misconducted occured and the case was essentially thrown out by the judge before further damage was done.

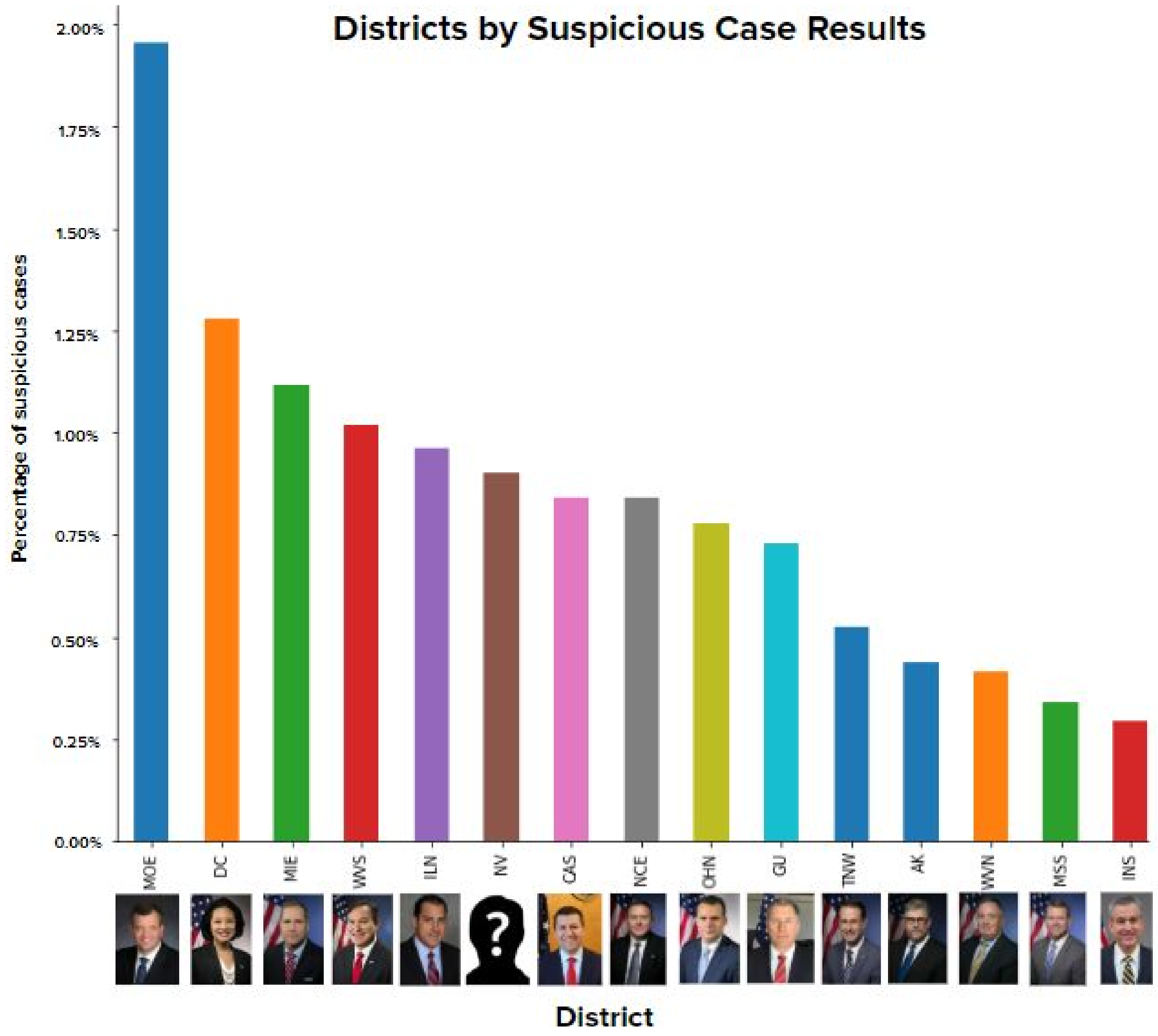
After identifying a list of possible indicators of misconduct, we eventually had our training labels list with either a 0 for no misconduct or a 1 for misconduct, and these labels were only for our dataframe where we actually had a disposition reason. The rows of data that did not would end up as our data that we would classify ourselves to determine if there might have been misconduct based on our classification models. To classify the training dataframe, we dropped several irrelevant features from the dataframe and used the built in pandas get_dummies() method which would transform all our data into numerical values. We finally were able to train our classifiers so that we could use them on our dataset in which we did not have a disposition reason to determine misconduct or not.

We attempted two different classification techniques - decision trees and logistic regression. After being fitted to our training data, both ended up with a similarly high cross validation accuracy of about 99%. We suspect this is due to a lack of available features in our dataset, which we'll address later. Despite their similar performances, we ended up using the logistic regression model in our final analysis due to its probabilistic nature.

After finding out that vital information such as the names of the prosecutors were redacted (or just missing), we were limited on the amount analysis we could do. Thus, we pivoted our analysis on looking at where misconduct occurs with regards to specific categories. We simply appended our predicted misconduct labels onto our testing dataset to create a dataframe with every single case with information such as District, Lead Charge, Judge name, etc. After, we filtered out the rows in which the misconduct label was a 0, so that we were only looking at cases in which it was predicted that misconduct was involved. Then we were able to use the value_counts() method to see which specific value of these categories was most present regarding a misconduct.

Results

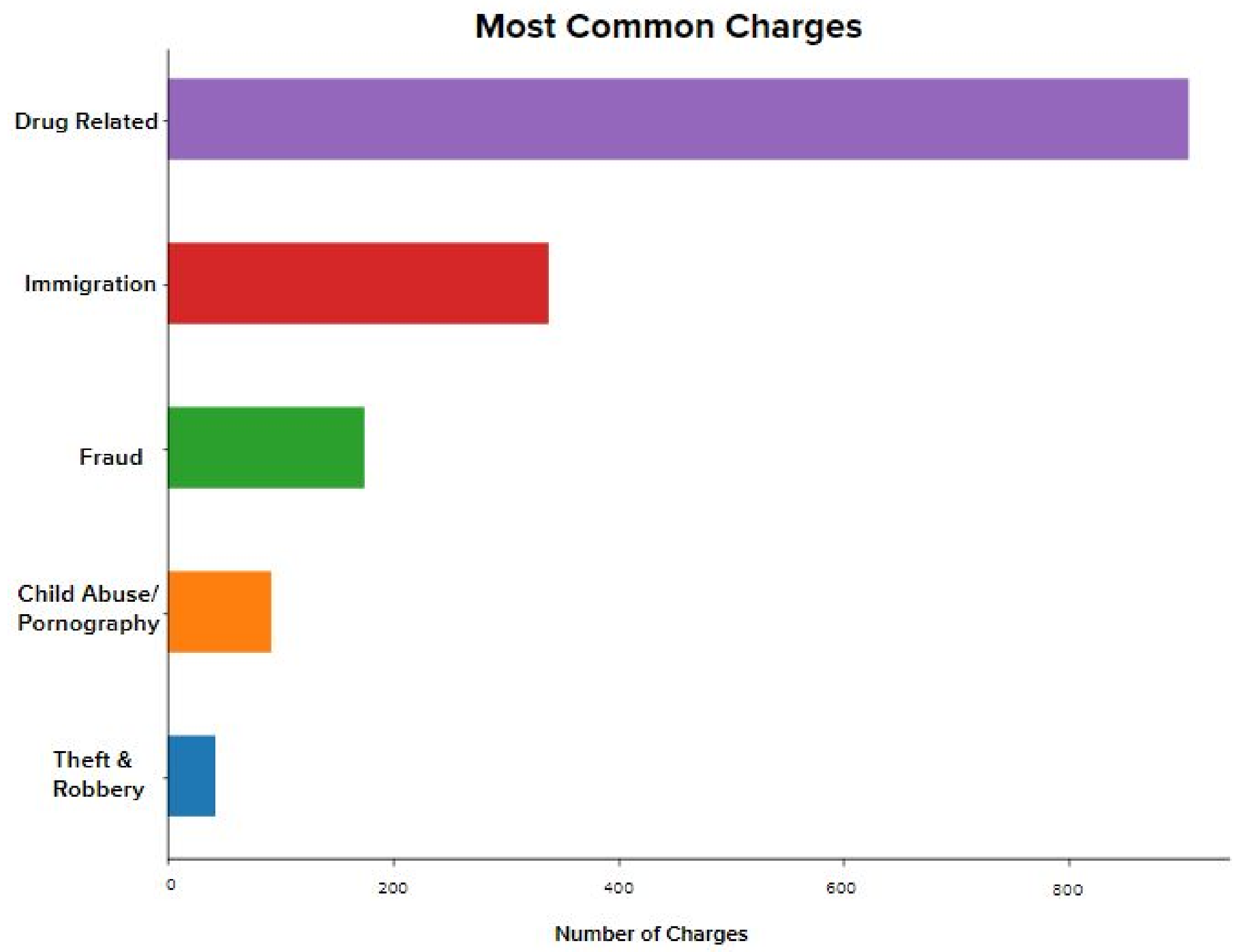
The first category we analyzed was looking at which district shows up the most when it comes to misconduct. Instead of doing a total count, we decided to use a percentage of total cases that might have experienced misconduct so that the results were not skewed and large districts such as Southern California did not automatically show up as a top problem area (incidentally, the southern district of California did show up as a problem area even when using percentages). The resulting graph is shown below:



1. **MOE** - Missouri Eastern - Jeffrey Jansen
2. **DC** - District of Columbia - Jessie K. Liu
3. **MIE** - Michigan Eastern - Matthew Schneider
4. **WVS** - West Virginia Southern - Matt Stuart
5. **ILN** - Illinois Northern - John R. Lausch, Jr.
6. **NV** - Nevada - Dayle Elieson?
7. **CAS** - California Southern - Adam L. Braverman
8. **NCE** - North Carolina Eastern - Robert J. Higdon, Jr.
9. **OHN** - Ohio Northern - Justin E. Herdman
10. **GU** - Guam - Shawn N. Anderson
11. **TNW** - Tennessee Western - D. Michael Dunavant

12. **AK** - Alaska - Bryan Schroder
13. **WVN** - West Virginia Northern - William J. Powell
14. **MSS** - Mississippi Southern - Mike Hurst
15. **INS** - Indiana Southern - Josh Minkler

The results seem to be pretty well distributed across the U.S., geographically speaking, with several districts from the midwest, west coast, and south appearing. Next, we looked at what was the most common lead charge in cases where misconduct was likely to have occurred. Here was our resulting graph:



The results above are striking. Drugs and immigration far outweigh all other categories. This was exactly the suspicion of our partner, Brooke, when it came to cases that involved misconduct. There appears to be an epidemic of cases in the US involving these two charges where either through unlawful evidence or a biased system people are being accused, but eventually the case gets thrown out or brought to appeal when it is found out there was misconduct.

Our results, however, are not perfect. Due to much of the data being redacted, we were only able to perform the training of our classifiers on a limited number of features. This is made clear in the validation accuracy of our data, which sits bizarrely high at over 95%. For something as murky as the criminal justice system, one would expect a little more variance in our results. As such, the data certainly deserves a more robust investigation.

Future Work

Though we recreated the LIONS database, much of the data in it is scrubbed at the discretion of the Department of Justice. Using our classifier, it could be possible to guess the outcome of these cases when going deeper on the database. In addition, if we could access the complete version of file “gs_archive_case.txt”, we can get the names of the prosecutors that are overseeing the cases, which could be a useful lead into misconduct in the Department of Justice, as this is currently missing from the dataset.