**discovering the real value of buying a politician**

ANLY-501 & COSC-587, Project Assignment 1

October 16, 2015

Tim Ahn

Arif Ali

John Hotchkiss

Joshua Kaplan

Hongkai Wu

The power of a U.S. congressman lies in the influence they wield to shape the identity of the country. In many cases, a candidate’s ability to raise funds for a campaign will be the deciding factor in getting elected to Congress. This critical fact creates an environment of careful platform planning and strategic alliances, quite often with specific industries. Data representing the contributions every candidate has received and final election figures are publicly available through the Federal Election Commission (FEC). Also available, through various public channels, are daily stock quotes and financial metrics for all publicly traded companies. But what is unknown within these data exists in the wake of the final votes and beyond. The ambiguity of the actual impact of the moneys raised to the contributing industries perpetuates the current process as the status quo. Are these companies’ bottom lines truly benefitting from the financial support they provide to candidates or is the idea of buying influence simply recognizable hope without an actual realized gain? Manufacturing new data from the existing, seemingly unrelated, election and financial resources to provide a measurable outcome of the impact that political contributions have on an industry will finally allow us the opportunity to assess their returns.

The data being collected for this study will begin with a breakdown of political contributions by industry for each general election candidate. They will include the top industries with the highest amount of contributions for each candidate. This allows us to determine the source of contributions and to whom they are going. Also being collected are recent congressional election results for each Representative district and Senate seat. These data will include candidate names, state and district, party affiliations, votes, and percentage of total votes collected. This information can be used to identify clusters related to geographic location, the effectiveness of political contributions, and industry concentration within the population, among other correlations. Finally, we will collect historical financial data of relevant industries (those that financed politicians, and can therefore be connected to them) within a major stock market index, as well as industry-specific indices to provide a fair representation of an industry’s overall performance. Included are daily prices for open, close, high, low, total volume, and adjusted close.

The goal of this study will be to uncover possible relationships between political contributions by industries to specific parties and candidates and the successive impact to the market as a whole and to itself. The expectation is to see a significant correlation between the industry-wide contributions to a candidate and that particular candidate’s likelihood to win an election. Furthermore, we expect to observe a substantial benefit when an industry supports an elected candidate. It may also be possible to explore the magnitude of correlation between industry spending and actual vote counts within a district and/or state. Another intriguing consideration will be industry contributions compared to the market value of the entire industry and comparisons of this measurement between industries.

The following provides an overview of each data source in use for our study and any related data issues that have been encountered.

**Political Contributions**

**Description**

We will collect data on contributions to congressional candidates from the 2004-2014 general elections, separated by industry. Data for 2012 and 2014, for election winners, will be obtained using Opensecrets’ get Legislators and candIndustry APIs. Data for 2004-2014 election winners and loserswill be obtained by scraping Opensecrets.org. Opensecrets gets this data from the FEC. Both the API and scraping were used since the API provided more attributes per candidate-year combination data than could be scraped.

**Issues**

* Missing data on some candidates from 2012-2014 (6 total) in the API data due to a glitch in Opensecrets’ data collection – they have been notified of the glitch but fixes have not yet been made.
* Some of the names in the scraped data did not extract cleanly – names with accents presented an encoding problem when brought into Python. This may cause issues when we merge our data.
* The scraped data includes all industries from which a candidate received at least $200 in a single contribution, while the data from the API only includes the top 10 industries that donated to the candidate, measured by aggregate contributions (ties at 10th contributor retained).

**Election Results**

**Description**

Election outcomes (vote counts and percentages) for every House of Representatives and Senate race from 2004-2014. Data for 2004-2012 will be downloaded from FEC.gov as csv files, and 2014 data will be scraped from NYtimes.com.

**Issues**

FEC 2004-2012 Election data (retrieved from <http://www.fec.gov/pubrec/electionresults.shtml>):

* FEC does not have 2014 election data available on their website, so we will use a separate source (NY Times 2014 election data).
* No vote count is available for unopposed elections.
* Data will have to be heavily manipulated in order to stack with the NY Times dataset and subsequently merged onto the main data set.

NY Times 2014 election data (retrieved from http://elections.nytimes.com/2014/):

* No data are available for uncontested elections – while the candidate may have received nearly 100% of the vote, it might be useful to know the exact percentage as well as the number of votes they received.
* Some of the names in the scraped data did not extract cleanly – names with accents presented an encoding problem when brought into Python.
* Data will have to be reshaped prior to incorporation with the primary dataset. Rows in this dataset contain information by state-year-district-race (multiple candidates per row) rather than state-year-district-candidate.

**Financial Market Data**

**Description**

We will collect daily stock market data for the S&P 500 and Dow Jones Industrial Average from 2004 to 2014 from Quandl.com. In addition, we will collect data on 5 sector-specific indices: the Nasdaq Banking Index, Nasdaq Industrial Index, Nasdaq Biotechnology Index, NYSE Amex Oil Index, and the PHLX Gold/Silver Index. S&P and DJIA data will be obtained using the API, while data for the sector-specific indices will be downloaded as csv files.

**Issues**

* S&P 500: Missing data for Navient Corporation (ticker NAVI); the API did not work for this ticker, as the page for Navient apparently does not exist on Quandl.com. The rest of the data are available, and did not contain any noise.
* Dow Jones: Same source as S&P 500, no noise or missing data

**Sector Specific Indices**

Nasdaq Banking Index:

No issues; downloaded csv from <https://www.quandl.com/data/NASDAQOMX/BANK-NASDAQ-Bank-BANK>

*The NASDAQ Bank Index contains securities of NASDAQ-listed companies classified according to the Industry Classification Benchmark as Banks. They include banks providing a broad range of financial services, including retail banking, loans and money transmissions. On February 5, 1971, the NASDAQ Bank Index began with a base of 100.00.* (from Quandl)

Nasdaq Industrial Index:

No issues; downloaded csv from

<https://www.quandl.com/data/NASDAQOMX/INDS-NASDAQ-Industrial-INDS>

*The NASDAQ Industrial Index contains securities of NASDAQ-listed companies not classified in one of the NASDAQ sector indexes. These include firms that are involved in oil and gas productions, oil equipment, services & distribution, chemicals, forestry and paper, industrial metals, mining, construction and materials, aerospace and defense, general industrials, electronic and electrical equipment, industrial engineering, support services, automobiles and parts, beverages, food producers, household goods, leisure goods, personal goods, tobacco, food and drug retailers, general retailers, media, gambling, hotels, recreational services, restaurants and bars, travel & tourism, electricity, gas distribution, water, and multi-utilities. On February 5, 1971, the NASDAQ Industrial Index began with a base of 100.00.* (from Quandl)

Nasdaq Biotechnology Index:

No issues; downloaded csv from

<https://www.quandl.com/data/NASDAQOMX/NBI-NASDAQ-Biotechnology-NBI>

*The NASDAQ Biotechnology Index contains securities of NASDAQ-listed companies classified according to the Industry Classification Benchmark as either Biotechnology or Pharmaceuticals which also meet other eligibility criteria. The NASDAQ Biotechnology Index is calculated under a modified capitalization-weighted methodology. The Index began on November 1, 1993 at a base value of 200.00.* (from Quandl)

NYSE Amex Oil Index:

No issues; downloaded csv from

<https://www.quandl.com/data/YAHOO/INDEX_XOI-NYSE-AMEX-Oil-Index>

*The NYSE Arca Oil Index, previously*[***AMEX***](https://en.wikipedia.org/wiki/American_Stock_Exchange)***Oil Index****,*[*ticker symbol*](https://en.wikipedia.org/wiki/Ticker_symbol)*XOI, is a price-weighted index of the leading companies involved in the exploration, production, and development of petroleum. It measures the performance of the oil industry through changes in the sum of the prices of component stocks. The index was developed with a base level of 125 as of August 27, 1984.* (from Wikipedia)

PHLX Gold/Silver Sector Index:

No issues; downloaded csv from <https://www.quandl.com/data/YAHOO/INDEX_XAU-PHLX-Gold-Silver-Sector-Index>

*The PHLX Gold/Silver Sector Index (XAU) is a capitalization-weighted index composed of companies involved in the gold or silver mining industry. The Index began on January 19, 1979 at a base value of 100.00; options commenced trading on December 19, 1983.* (from <https://indexes.nasdaqomx.com/Index/Overview/XAU)>

**Data Cleaning**

In order to properly score the data, we evaluate two categories of data issues. The first is potential outliers in the quantitative attributes within our data. The second is a verification of completeness and formatting.

A confidence interval of 99% is used to determine outliers. That involves looking for any quantitative values that are greater than or less than three standard deviations away from the mean of the attribute. The percentage of values outside this confidence interval is recorded for each attribute. Finally, the outlier percentages are averaged together across the data set to generate a final outlier score.

The second component is slightly more involved and requires two subcomponents. The first subcomponent requires checking the percentage of missing values for a given attribute. Thus the total number of blank values in the attributes were added up and then added to the second subcomponent. This subcomponent was designed specifically for the string attributes. The string attributes are important for merging; thus basic formatting needs to be standard. For the Names in particular, we evaluated whether or not they had a proper case. To do this, we split up the strings by word and counted the number of words, thencompared the number of words to the number of upper case letters in the string. For all strings in an attribute that did not follow this rule,match it were summed of to make up the second subcomponent of the second component.

After creating the two sub-scores above, they are weighted equally and subtracted from 1 , yielding one final percentage. After running through each of the data sets, we are able to make general conclusions about the data. For data originating from an API, the percentage is significantly higher than that of data originating from scraping or csv downloads. The worse scoring data set was scraped from the New York Times. This indicates a flaw in the scoring algorithm since in this file missingness might be totally fine on an attribute by attribute basis. Also, due to the way it was scraped, some numeric attributes remained as strings. In order to compensate for the string form of those attributes, the data cleaning script performed light editing to convert them to a numeric format. There were two unexpected results of the cleaning algorithm. Since the FEC data was obtained in Excel Format and theoretically preprocessedit was expected to score significantly better than the scraped New York Times Data Set, however it was a surprise that it did so much worse than the API obtained Data sets. This is probably due to the FEC focusing more on what appears to be a human-readable processing than a machine-readable processing. The second surprise was that, although the NY Times and some of the Open Secrets data were both scraped, the open secrets data had a much higher score. This is probably due to the difference in formatting, where the NY Times file would be considered a “wide” file and the Open Secrets data a “long” file.

As discussed briefly above, a possible drawback to the cleaning method we employed is that the program is blind to whether or not missing values are errors for a particular attribute. It could be in place of a zero, meaning there was no actual information gained. The proper case-base component could also be fooled by the data. Some common words, such as ‘for’, generally aren’t required to follow proper case, but will still be flagged by the program as errors. Luckily, this was not a problem with the names we will use to merge, which was our primary concern. However, that is not to say that the names don’t trip up proper case. The name Bill O’Reilly is flagged as unclean data because the name is two words, yet there are three upper case letters. Thankfully, the percentage of names impacted by this issue is low, therefore we should be able to verify merges for these cases with relative ease.