They Spent HOW MUCH??? Tracking Presidential Campaign Donations and "Independent Spending" across Time

W205 Project, Spring 2016 - Winston Lin and Alfred Arsenault

Project Description:

Although information about donations to political campaigns, candidate spending, and spending by "independent groups" in support of or in opposition to candidates for Federal offices is publicly available, many people do not know or access the information. This leads to rumors about who is really behind which candidates.

An example of the situation is Ohio's 8th Congressional District. That district was represented for over 24 years by former Speaker John Boehner, who resigned earlier this year. There is a special election to fill out the rest of his term, with the primary to be held in March. There are 15 Republicans running to replace him. To date, there have been 42 separate "independent" expenditures supporting or opposing one of the candidates, with the total amount spent exceeding $1.8 million dollars! Almost two million dollars spent in the primary of an election to fill a House seat for 8 months, in a race almost guaranteed to be won by a Republican due to demographics.

We plan to address this information problem by providing a system that can ingest authoritative data from the Federal Election Commission, process it, and make it easy to understand with our analysis tools.

Data Sources:

The primary data source will be the Federal Election Commission's Data Catalog, at http://fec.gov/data/DataCatalog.do. We will focus on two main data sets: (1) Candidate Summary (CS) and (2) Independent Expenditures (IE). The former will helps us examine direct donations to presidential campaigns, and the latter will shed light on third-party expenditures (advertisements, canvassing efforts) in support of (or against) a given candidate. The CS data includes itemized contributions from individual contributors and "other committees" (PACs, various special interest groups), as well as transfers from authorized committees. The IE data includes itemized expenditures made by third parties in support of (or against) a given candidate.

Files are available in comma-separated value format. Initially, we will only use files for the 2016 presidential election cycle in order to build the full-solution. Once the "steel thread" is completed, we may incorporate contributions for past presidential election cycles, past presidential candidates, and candidates for the House of Representatives and Senate.

Volume of data. We inspected itemized contributions given to Hillary Clinton for the 2 years leading up to her 2016 and 2008 presidential campaign. The total volume of data was 112MB, which we will assume is an upper bound for the volume of data for a single candidate, as she is a leading

Democratic candidate and is running for President a second time. Our project will initially examine all remaining 2016 presidential candidates from the Republican and Democratic parties (6 in total), so the total volume of data involved in our project should be no more than 1GB.[1]

Variety of data. The CSV files containing itemized contributions (CS data) for each candidate all have similar schemas that include fields for contributor profile information, details of each contribution, and other metadata for the recipient and transaction. The file for individual contributions has two additional fields for employer and occupation that are not present in the files for contributions from other committees and transfers from approved committees. The schema for files containing independent expenditures (IE data) have a different schema that includes fields for third-party profile information, details of each expenditure, and other metadata for the targeted candidate and transaction.

Velocity of data. Candidate reports must be filed quarterly, but the FEC updates the candidate file nightly and provides an RSS feed to notify subscribers of developments in which they may be interested. We will access the Candidate file once per day, but we can notify users that the information may not be current depending on the date of the candidate's most recent filing. "Independent Expenditures" must be reported within 48 hours of the expense - 24 hours as the election draws close - and the FEC tries to push new entries out within minutes of receiving the reports. We will plan on accessing the "Independent Expenditures" files once per day to constrain network bandwidth.

Data Ingest and Storage:

We will explore how to best ingest the data. We are currently considering deploying an Apache Kafka cluster on AWS as we believe that this will give us the best reliability and interoperability solution. Once ingested and cleaned, the data will be dumped into AWS S3 buckets for storage. The data can be moved over to an AWS ESB attached to an EC2 instance when it is ready.

Reporting and Analysis:

Query engine. Results reporting will be done using a searchable SQL interface. There will be "pre-canned" reports that can be run to provide results we expect to be requested often - e.g., "how much do Goldman Sachs employees give to which candidate?" The users will be provided with the SQL interface so that they can create whatever additional queries they want.

Visualization. We will also provide a Tableau workbook as a visualization layer to give non-technical users a click-and-drag interface for ad-hoc queries. This will also enable easier analysis and interpretation of data. Tableau will allow our aggregate contribution and expenditure data, as well as any derivatives such as segmentation

[1] "Who Is Running for President?"
http://www.nytimes.com/interactive/2016/us/elections/2016-presidential-candidates.html?_r=0