# Assuring Election Integrity: A Comprehensive Ecological Framework for Evaluating Elections in Southern California
# 2017-2019 Project Final Report

R. Michael Alvarez
California Institute of Technology

June 11, 2019

## Executive Summary

In 2018, we launched an ambitious pilot study to develop and implement a wide variety of tools to study the integrity of elections in Orange County (CA). In this report, we review our project, and provide a summary of the results from our analyses.

The election integrity methodologies that we developed, tested, and deployed in 2018 using data from Orange County included the following analytical applications:

- A methodology for scanning the County's voter registration database for anomalous changes;
- Tools for analyzing post-election precinct-by-precinct datasets for unusual patterns in voter turnout and votes cast;
- Monitoring of social media data for mention of the election and any issues that arose;
- In-person election observation in early and Election Day voting locations;
- Surveys of registered voters about their experiences with the registration and voting processes and technologies.

Across our major analytical studies, we produce evidence that the administration and conduct of the 2018 primary and general elections in Orange County had a high degree of integrity. Orange County is a very large election jurisdiction, with complex and evolving procedures. That our analytical tools found so few areas of concern lends strong confidence in the conclusion that the county's election administration was without compromise.

In this report, we also present our plans for scaling this election integrity analysis to a larger set of Southern California counties in 2020.

## Assuring Election Integrity

With the generous support of the John Randolph Haynes and Dora Haynes Foundation, through a research grant to the California Institute of Technology, we conducted an ambitious pilot project to develop, test, and implement an ecological framework for evaluating election integrity. Working with the Orange County Registrar of Voters (OCROV) in Southern California, our team produced an array of different methodologies for this comprehensive analysis, using data from the 2018 primary and general elections in Orange County. Our analyses looked at data on early voting, voting by mail, Election Day voting, turnout, and voter registration; the data we used came from administrative sources, voter and poll worker surveys, in-person election observation, post-election audits, and social media.

The primary objective of this project was to demonstrate how to collect, analyze, and disseminate these various analyses, in a public and timely way. By conducting these analyses, we were able to provide important feedback to the OCROV, to voters, and to other stakeholders, helping to bolster confidence in Orange County's 2018 elections. In this report, we outline the project's goals and achievements, and our plans for the future.

## Project Summary and Achievements

This project kicked off in April 2018, just before the 2018 statewide primary elections in Orange County. The project had three overarching objectives:

1. Develop and implement a wide array of overlapping election administration and quantitative methods that would together provide a holistic and ecological perspective on the integrity of the 2018 primary and general elections in Orange County;

2. Provide timely and useful reporting to the OCROV regarding the results of these analytical tools;

3. Disseminate the results of our analyses quickly in a public and transparent manner, to assure voter and stakeholder confidence in the integrity of the primary and general elections in the county.

The initial phase of the project involved the establishment of a public website for the dissemination of our analyses, which we accumulated into primary and general election dashboards. The website, https://monitoringtheelection.us, was built to distribute the analyses, display our dashboards, and provide other information about the project.

This initial phase of the project also involved establishing a strong collaborative working relationship with the OCROV. Our team required access to information and data for the project, and we also benefited greatly from the expertise and knowledge base at the OCROV concerning how elections are conducted in the county. Immediately after our project launched, we met with OCROV a number of times to establish a strong working collaboration at the project's inception, allowing us to be in communication with the OCROV prior to the June primary.

For the June 2018 primary, we focused much of our attention on the development of a sophisticated methodology to evaluate the security and integrity of the county's voter registration database. This required that our team to learn about the structure of the OCROV voter file, how the OCROV manages the database, and the best way to create a secure data pipeline. Afterwards, our team worked to build a series of software applications that scan daily voter file "snapshots" for duplicates, record deletions, record additions, and changes to information within these records. After we built a time series of these metrics, we were able to build an application that detects when a daily rate of change in the voter database represents a statistical anomaly worthy of further examination by the OCROV. The voter registration audit began in May 2018, and is still ongoing.

Another ongoing analytical tool that we launched in May 2018 is our Twitter monitoring application. This application constantly queries the Twitter Streaming API for tweets about various election issues. For all relevant tweets, our algorithm proceeds to download, pre-process, geo-tag, and store all data. Throughout the 2018 election cycle, our website displayed metrics on all tweets for a variety of election keywords; we have also analyzed the subsample of tweets we could geolocate to Orange County, and more generally in California, to determine the utility of this methodology in analyzing state and local election integrity.

Specifically for the June 2018 primary, we implemented a series of additional election integrity studies:

- In-person observation studies of early voting at OC vote centers.
- In-person observation studies of Election Day voting.
- Statistical forensic analyses of precinct turnout.
- Statistical forensic analyses of candidate vote shares from the Statement of Votes.
- Observation of post-election risk-limiting audit.
- Examination of OCROV poll worker survey results.

For the November 2018 general election, we utilized these same set of methodologies; the only major difference in the general election was the implementation of two different types of voter experience surveys. We conducted one of the voter experience surveys using contact information from registered voters in the Orange County voter file, and we conducted the other in parallel with a related study by researchers from the University of California, Irvine.

# Key Findings

## Voter registration auditing

We began our voter registration database auditing analysis on April 26, 2018, and have continued to monitor the OCROV voter registration database since that date. Our registration auditing approach looks for changes in daily snapshots of the voter file: new records that have been added, previously existing records that have been dropped, and records where important fields (for example, name, registration address, date of birth, or party registration) may have been changed. Our methodology compiles a database of those changes since April 26, 2018, and then we use statistical anomaly detection to determine when a particular day exhibits changes which are outliers and thus merit further examination. When our script detected anomalous events, our team communicated those findings to the OCROV. Periodically, we posted summary reports on the project website.[1]

Our last post-2018 general election report on our voter registration auditing was conducted on December 12, 2018. At that time, we noted that "our algorithm and interquartile range (IQR) analysis have found 36 events. These events all appear to be the result of normal database maintenance activities by OCROV." In the time since we wrote that report, we have continued to monitor the OC voter registration database, and we have not detected any unusual or anomalous events.

## Voter experiences

In the 2018 pilot, we involved both qualitative and quantitative approaches for studying voter experiences with election administration and voting technologies in Orange County.

The qualitative approach used in-person election observation. In both the primary and general elections, we deployed small teams of trained election observers to early voting and Election Day voting locations. Observers collected information on each location (for example, by looking at the availability of parking, whether the voting location was easily accessible, and making sure that signage was available to make clear the voting location's exact location and entrance), on the layout of the voting location, lines and voter wait times, and on any other issues of note during their visit to the voting location.

These qualitative studies yielded a wealth of useful information, for both our research team and the OCROV. We were able to provide a great deal of feedback to the OCROV from these studies about the potential quality of the voting experience in the locations our observation teams visited. We produced detailed reports that we published on the project website, detailing our observations and making recommendations for improvements. Our teams were able to visit most of the early voting locations, but only a small fraction of Election Day polling places. Polling places were selected for inclusion in these studies using a variety of factors: we generally selected polling places that were

[1] See `https://static1.squarespace.com/static/5ace8a6b45776eba2e40cbee/t/5c12d7f10e2e724784d6833a/1544738801989/VR_Audit_Report.pdf` for an example.

in competitive U.S. House districts, and which represented the many different types of locations used for Election Day voting. Overall, our observing teams did not find significant issues at the observed voting locations; our findings indicated that most voters were able to easily and securely cast their ballots in person or drop off ballots received by mail.

We also built and deployed two large-scale post-election voter experience surveys. Our primary voter experience survey contacted Orange County registered voters using email addresses from the voter registration database, and invited them to participate in our survey. We received 6,948 completed responses in November 2018. This voter experience survey produced two important topline results:

- We found that 87% of voters responding to this survey were confident that their vote was counted as they intended; 86% indicated that they were confident that votes in Orange County were counted as intended.

- Nearly all of the responding registered voters reported a positive voting experience, whether they voted by mail, in an early voting location, or on Election Day: 96% of responding voters said that it was very or fairly easy to find their polling place; 95% of by-mail voters did not have any trouble getting their ballot by mail, 97% did not report any trouble making their ballot.

We also conducted an online survey of Orange County residents as part of a larger survey study conducted in 2018 with other researchers. As we use this online survey to conduct other research projects in 2019 and 2020, we will continue to post any findings relevant to the integrity of the 2018 on our project's website.

In conclusion, our in-person election observation studies and voter experience surveys together indicate that in the 2018 primary and general elections, there were very few problems for vote-by-mail, early, and Election Day voters. Our analyses indicate that most Orange County voters have had good experiences voting and maintain strong confidence in the integrity of the 2018 elections their county, impressive given the larger context of an election environment that was quite competitive, had strong voter turnout, and took place in a very large election jurisdiction.

## Forensics

In 2018, we conducted two types of forensic analyses using Orange County data, in both the primary and general elections, both using post-election public reports on precinct-by-precinct results from the primary and general elections.

The first type of statistical forensic analysis that we utilized for this project involved examining precinct-by-precinct post-election reports of voter turnout. The OCROV posts precinct-level voter turnout PDF reports frequently in the post-election period. We built software scripts that accessed these PDF files and scraped the precinct-by-precinct voter turnout data. After preprocessing these data, we examined them graphically. Previous research notes that graphical representations of the

distribution of the percentage of turnout in an election across precincts should typically be unimodal and should have an approximately normal distribution (i.e., the distribution of the percentage of voters turning out across precincts should look like a "bell-shaped" distribution).

In our forensic analyses of the OCROV voter turnout reports, we consistently found that the distributions were unimodal and approximately normal — the type of forensic evidence that indicates very few anomalies. These forensic studies did find a very small number of anomalies in the data, which upon further investigation by our team and by the OCROV were determined to be the result of reporting or administrative errors in precincts on Election Day. These anomalies were found in only one or two of the more than 1,500 precincts used in Orange County in the 2018 primary (1,561 precincts) and general (1,546 precincts) elections, a very low incidence rate.

Secondly, we undertook various forensic analyses of the post-election "Statement of Votes" (SOV) provided by the OCROV on their website during the post-primary and post-general election period. The OCROV periodically posts these reports in PDF files during their canvass period; we built a script that accesses these files and scrapes the relevant data. We then constructed precinct-by-precinct datasets of votes cast for every candidate in each of the elections on the ballot. Using these datasets, we produced a variety of graphical and statistical forensic studies of the SOV data after the primary and general elections.

The initial SOV forensic reports focused on visualizations of the data from the top-of-the-ballot statewide candidate races, examining precinct-level histograms of candidate vote shares and scatter-plots of candidate vote shares by turnout. As with our turnout forensics, these visualizations look for anomalous distributions in candidate vote shares and in their relationship with turnout patterns. For example, seeing visualizations of candidate vote shares with multi-modality, or with highly skewed distributions, can be an indicator of issues that might need to be further examined. Seeing anomalous correlations between candidate vote shares and turnout, across precincts, could also be an indication of issues worthy of further examination. We found no evidence in these studies of significant outliers or unusual distributions that would merit additional investigation.

Following the 2018 general election, we undertook a more sophisticated statistical analysis of split-ticket patterns in the 2018 general elections in Orange County. Ticket-splitting occurs when a voter selects a candidate from one party in a race, but then selects another party in another race. For example, if a voter selected a Republican candidate for governor in the 2018 general election, and a Democratic candidate for the U.S. House of Representatives election, that is split-ticket voting.

Questions were raised about split-ticket voting in the competitive U.S. House elections in Orange County. Given that in past elections, Orange County has typically been considered a Republican stronghold, some observers of the 2018 election questioned whether Orange County voters were casting split-ticket ballots in the 2018 general election, or if the Democratic candidate successes in some of the competitive U.S. House of Representatives elections in Orange County anomalous. This led us to examined SOV data from Orange County (and from Los Angeles County for comparison) in order to determine whether ticket splitting was an anomaly worthy of further examination. We found nothing anomalous in our forensic analysis: ticket splitting in Orange County was widespread

(not isolated to specific House elections) and was quite similar to ticket splitting seen in neighboring Los Angeles County. Thus, these forensic studies showed no unusual anomalies requiring further investigation.

## Social media monitoring

We used social media monitoring to provide analytical data to study election administration and technology in Orange County. We built software tools that access the Twitter Streaming API and collect tweets that contain one or more tracked keywords. All of the tweets that the Streaming API passes to us are stored in a database, and then with other software tools we preprocess and then analyze the tweets. For this pilot project, we collected tweets that contained keywords associated with Election Day voting, remote voting, voter identification, polling places, and election fraud. This monitor ran continuously from August 5, 2018 through December 12, 2018, and stored nearly 29 million tweets from over 3.6 million unique users.

Metrics from the collected tweets were displayed on the project's website in real time. This allowed us to provide a public analysis of the conversation on Twitter about election administration and technology as the 2018 general election progressed. However, as far as election issues in Orange County are concerned, our Twitter election monitor does not provide any additional evidence of administrative or technological problems; our analysis of the subsample of tweets that we could link to Orange County (CA) shows that the volume of discussion about election issues in the county was lower than it was for California or the nation. This additional analytical perspective provides no evidence for election issues or anomalies needing further investigation in Orange County.

## Putting it all together

Based on our examination of our work in 2018, we believe that the project was a success. Methodologically, we developed and implemented a variety of different qualitative and quantitative tools that measured the integrity of different (and overlapping) aspects of election administration and technology in Orange County's 2018 election cycle. We were able to provide valuable feedback and analytical intelligence to OCROV in a timely manner, and to provide public reporting of our analyses and conclusions.

Substantively, our analyses all reach the same set of conclusions. Across our major analytical studies, we produce evidence that the administration and conduct of the 2018 primary and general elections in Orange County had a high degree of integrity. There were very few problems seen in any of our analysis, and given that Orange County is a very large election jurisdiction, with complex and evolving procedures, the fact that our analytical tools found so few issues provides strong confidence in the integrity of election administration in Orange County.

# Pivoting to 2020

Building off the success of the 2018 pilot project, we are currently in the initial phase of scaling this project to include other Southern California counties. We will continue studying Orange County in the 2020 election cycle, working to monitor the integrity of their voter registration data, continuing to monitor social media and voter experiences, conduct election forensics, and examine various potential concerns associated with the county's transition in 2020 to the use of new voting systems, voting by mail, and vote centers.

In 2020, we are expanding our study to include Los Angeles County, and the correspondence with Los Angeles county has been initiated, now building a secure data pipeline as we have in Orange County. We will be engaging in the same election analytical studies as in Orange County, examining the integrity of the county's voter registration database, studying voter experiences and social media, and undertaking forensic analyses after the March and November 2020 elections. We will also be working to help the county examine its transition to new voting technologies, vote centers, and the use of new types of ballot design in 2020.

We are also planning to reach out to other counties in the Southern California region to engage them in our election integrity study. We have begun working with Los Angeles County, and they have committed to working with us in 2019 and 2020. We hope to have the counties in Southern California as participants in our project in 2020.

Analytically, we are working on a number of improvements to the tools we have developed.

- We continue to work on improving our voter registration auditing applications, in particular, on building a model-based approach for statistical anomaly detection for 2020 that will reduce the number of false positives.

- We are working to improve our voter experience survey methodology for 2020, especially working to adjust our questionnaire to better measure some of the changes occurring in election administration in Southern California. We are also working on other methodological improvements that we believe will increase our survey's response rate.

- We are refining our social media applications, with the goal of enabling faster and more accurate geo-location of a larger proportion of the social media posts collected.

- We are examining our applications that gather, pre-process, and analyze post-election reports so that we can more quickly and effectively conduct statistical election forensics. We also will be working on the use of machine learning for election forensics.

- We are revising our project website and election integrity dashboards for efficient and effective communications of the project's results.

# Dissemination and Impact

Since the project was launched, we have presented preliminary results from the pilot project at a number of academic and research conferences:

- Southern California Methods Conference, UCLA, September 2018.
- Election Audit Summit, MIT, December 2018.
- Midwest Political Science Association Annual Meetings, Chicago (IL), April 2019.

We are planning on presenting results from the pilot project at other academic and research conferences in the summer of 2019, including the Election Sciences, Reform, & Administration Conference, 2019 Summer Meetings of the Society for Political Methodology, and the 2019 Pre-APSA Workshop on Securing Elections.

We are planning on a number of different venues for publication of our research as part of this pilot project. First, we have written a paper on the voter registration auditing methodology that is now under review at a peer-reviewed academic journal; we hope to have this paper accepted for publication by the end of the summer 2019. We are also working on papers that utilized data from the general election voter experience surveys, and we plan to have these papers under review by the end of 2019.

Most importantly, we are working on a book monograph, reporting on the primary quantitative approaches that we used in this project. That monograph, tentatively titled, "Securing American Elections: How Data-Driven Election Monitoring Can Improve Our Democracy," should be under review at a major university press by the end of June 2019.

The results of our pilot project were also disseminated in various ways to the interested public. The project's website has had 1,879 visits since project launch, 1,628 unique visitors, and 33,374 page views. We provided periodic updates about important project milestones on the project's website blog, as well as on the Election Updates blog. The project was written about in the Caltech Magazine, in Bloomberg News (*California Taps Data Scientists in Election Monitoring Bid*, June 5, 2018), the Pasadena Star-News, and the OC Weekly. It was also featured in Orange County's 2019 Vote Center Briefing Report.

The project was also used for teaching and curricular activities at Caltech. Three Caltech graduate students very heavily worked on the project. Two other Caltech graduate students were involved in some of the project's research activities. A number of Caltech undergraduates participated in election day observation activities in 2018; during the summer of 2018, one of the Caltech undergraduates worked closely with the project team as part of the Caltech Summer Undergraduate Research Fellowship program. Material and data from the pilot project has been incorporated into Professor Alvarez's courses, and some of the data collected by our Twitter monitor has been used by students working on independent research projects.

We will continue to maintain the project website through 2021. The final reports from 2018 will be

archived on the project website (we will start to highlight the new research that we are doing for 2020 more prominently on the project website). We have begu archiving code that we have built at the project's GitHub; when possible, we will also provide data on the project GitHub.[2]

# Conclusion

With the generous support of the John Randolph Haynes and Dora Haynes Foundation, we successfully conducted a study of the integrity of the 2018 elections in Orange County (CA). This study, which also required assistance and collaboration from the Orange County Registrar of Voters, analyzed the administration and technology of the 2018 primary and general elections in the County, allowing a holistic and ecological evaluation of the conduct of this competitive midterm election. We wish to thank Neal Kelley, the Orange County Registrar of Voters, and Justin Berardino (Operations Manager for OCROV) for their assistance.

We developed, tested, and deployed a number of innovative and unique methods (analytical voter registration data auditing, forensic analyses of precinct-by-precinct vote reports, and social media monitoring), along with other methods (in-person election observation and voter experience surveys). Examining the results of these analyses, independently and ecologically, leads us to conclude that the 2018 elections in Orange County were conducted with integrity.

As we are now moving quickly towards the 2020 presidential election cycle, we are working to improve our methodologies for use in the next election cycle. We will be implementing these improved methods for Orange County and Los Angeles County in 2020, and over the coming months will be working to recruit other Southern California counties to participate in this project.

[2] https://github.com/monitoringtheelection.

# Project Team

- Nicholas Adams-Cohen (Ph.D. Candidate, California Institute of Technology).
- Seo-young Silvia Kim (Ph.D. Candidate, California Institute of Technology).
- Yimeng Li (Ph.D. Candidate, California Institute of Technology).
- Spencer Schneider (Undergraduate Student, 2018 Summer Undergraduate Research Fellowship, California Institute of Technology).