Analysis of Electoral Fraud in the 2014 Afghan Presidential Election

Laksh Bhasin[1], Kurtis Carsch[1], Sharon Yang[1]
Instructor: Peter Ordeshook.
[1] *California Institute of Technology, 1200 E California Blvd., Pasadena, CA 91125, USA*

Abstract

This analysis searches for fingerprints of fraud in the second-round results of the 2014 Afghan Presidential Election. We look at various indicators in this search, such as turnout distributions; winning margin analyses; electorate share vs turnout plots; vote share vs turnout plots; correlations between observer deployment and turnout; and digit analysis. To pinpoint regions with possible fraud, we look at high-turnout districts and provinces with skewed vote-share distributions. These indicators suggest that both candidates – Abdullah Abdullah and Ashraf Ghani – are likely to have participated in noticeable amounts of regional ballot stuffing, with the winner (Ashraf Ghani) engaging in an overwhelming majority of the stuffing. Furthermore, our last-digit analysis implies that election officials falsified vote totals in many of the polling stations where Ghani won. In order to mitigate future cases of fraud, we propose increasing the number of observers present in the problematic provinces of Khost, Paktya, Paktika, Wardak, Parwan, and Panjshir, as well as increasing the number of available ballots at each polling station. Independent background checks on election officials might also be necessary, since several claims of bias were reported to the media. Lastly, we strongly implore the Afghan government to implement a census program and collect reliable and freely-accessible data on each district's population and number of eligible voters. These endeavors will aid future election audits, and will generally be useful for the developing country.

1 Background Information on Afghanistan

1.1 A Brief History

The geopolitical history of Afghanistan has been shaped by the heterogeneity of its diverse ethnic groups and leadership power struggles [1]. During the Chinese Han dynasty, the Silk Road passed through the country's mountainous terrain, connecting central Asia and the Indian subcontinent to Iran and the Western world. Although this trade route allowed communities in Afghanistan to thrive economically, this intermediate positioning was problematic; over the centuries, several destructive wars were waged by Alexander the Great (330 BCE), Ya'qub ibn al-Layth al-Saffar (870 CE), and Genghis Khan (1220 CE). With Ya'qub's conquest came Sunni Islam, which remains the most prominent religion in today's Afghanistan [2].

Afghanistan was subsequently ruled by the Safavids and the Mughals. In the 1700s, the Pashtun Ahmad Durrani conquered most of the territory that comprises today's Afghanistan. Durrani is widely regarded as the founder of the modern state of Afghanistan, and is also the subject of many Pashtun legends.

Conflicts with British and Russian forces led to the First and Second Afghan Wars during the 19th century. After the Third Afghan War in 1919, King Amanullah Khan established Afghanistan as a sovereign state free from British rule [3]. The nation was modernized through compulsory elementary education, the reduced influence of prominent Islamic figures, and new educational opportunities for women [4].

For several years, Afghanistan remained politically stable, until Pakistan-trained communist rebels seized government control in 1978. This later gave way in 1979 to a regime controlled by the communist Soviet Union, which helped strengthen the political rights of women.

The Soviet Union eventually withdrew in 1989, sparking several internal civil wars for political power. In 1996, the radical Taliban took control of the government and established the Islamic Emirate of Afghanistan. During their short-lived regime, the Taliban were notorious for their sexist, savage, and anti-Shiite views; in several instances, they orchestrated mass killings of Shia Muslims. Following the destruction of the Twin Towers in New York City on September 11, 2001, NATO forces launched an anti-Taliban campaign in Afghanistan. As a result of this campaign, the Taliban regime was removed by December 2001, and Hamid Karzai was installed as the head of state. In 2002, Karzai was appointed the nation's two-year interim president.

After the adoption of a new constitution, Karzai ran for president in 2004; his win made him the first democratically-elected leader in Afghan history. In 2009, Karzai was reelected as president for another five-year term, which ended in September 2014.

1.2 Government

The constitution of Afghanistan established the modern-day government in January 2004. This government is split up into a legislative branch, an executive branch, and a judicial branch [5].

The bicameral National Assembly of Afghanistan represents the country's legislature. This body is composed of two chambers: the House of the People (Wolesi Jirga) and the House of Elders (Meshrano Jirga). Elections for the former were first held in 2005, in which 250 candidates were elected at the district level to serve five-year terms. The latter house has 102 members; 34 of these members are elected by district councils (one per province) for three-year terms, another 34 are elected

by provincial councils (one per province) for four-year terms, and the remaining 34 are nominated by the president for five-year terms.

The executive branch is led by the president, first vice president, and second vice president. Both vice presidents run on a joint ticket with the president. This branch also includes a Cabinet of 25 ministers that supervise the laws proposed by the two Houses. Ministers are appointed by the president, but they must be approved by the National Assembly.

Following the 2014 presidential election, the post of "Chief Executive Officer (CEO)" was created to assist the president and lead weekly policy meetings with the Cabinet [6]. This position was created after a power-sharing agreement was signed between the front-running candidates (see Section 2.3).

1.3 Presidential Elections

The President of Afghanistan is elected every five years. The election is headed by the Independent Election Commission (IEC), which is designed to avoid bias from the current governing party. Candidates must announce their intent to run for the presidency a year in advance. To be elected, a single candidate must win at least 50% of the votes in a given election round. If no candidate achieves a majority in the preliminary election, a second runoff election is held between the two front-running candidates of the preliminary round. Each individual above 18 years of age is allowed to register for a single vote per election round.

To mitigate fraud, the EU and UN deploy observers for each of the major voting rounds. Both bodies also audit the results of each election when necessary, and help mediate contentious results. Furthermore, to ensure security and encourage citizens to vote, several guards are deployed at polling centers. However, despite this increased security during election season, numerous polling centers were closed during the runoff election due to threats from the Taliban [7].

1.4 Ethnicity Distribution

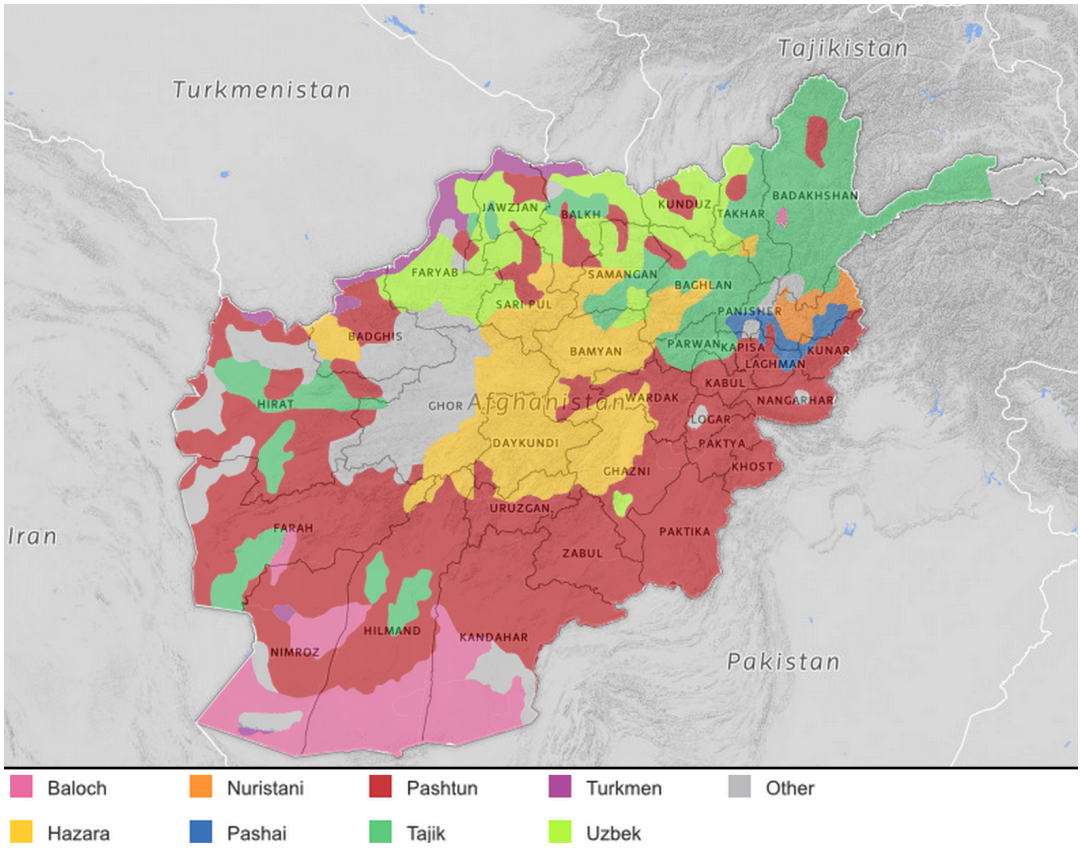


Figure 1: *Ethnicity distribution of Afghanistan. Pashtun is the most dominant ethnicity in the country. Regions with no coloring are sparsely populated. Figure modified from [8].*

The immense ethnic diversity of Afghanistan is an important factor in our analysis of election fraud. Due to historical ethnic tensions, many Afghans align with candidates of a similar background, instead of focusing as much on political ideologies.

The most prominent demographic of Afghanistan are the Pashtuns, who dominate the central, southern, and eastern portions of the country (see Figure 1). Consequently, the Pashtun ethnicity is more politically represented than other ethnicities in Afghanistan.

The front-running candidates of this election – Ashraf Ghani and Abdullah Abdullah – are both of Pashtun origin. Abdullah is also half-Tajik (a prominent ethnicity in the urban capital of Kabul), and this mix of ethnicities has alienated him in many Pashtun regions (relative to Ghani) [9]. The remaining nine candidates in the preliminary round were all of Pashtun background.

2 The 2014 Afghan Presidential Election

The 2014 Afghan presidential election marked a significant milestone in the political history of Afghanistan. In particular, this was the third presidential election since the fall of the Taliban in 2001, and also the first opportunity for a new leader in almost twelve years. The previous president, Hamid Karzai, was installed as an interim president in 2002 and elected in both the 2004 and 2009 presidential elections. After serving his second five-year term, he was constitutionally barred from rerunning [10].

Although the names of twenty-seven candidates were submitted to the IEC, only eleven of those candidates were eligible to run; the remaining sixteen were unable to provide sufficient documentation on their background. Three candidates subsequently withdrew to support Independent Centralist Zalmai Rassoul. In the run-up to the first round of voting, public opinion polls predicted that Abdullah Abdullah and Ashraf Ghani would each hold approximately 20-30% of the overall votes.

2.1 First-Round Results

The preliminary election was held on April 5, 2014. In this round, Abdullah won a plurality of 45% of the total votes, followed by Ghani with 31.6% and Zalmai Rassoul with 11.4%. The election was characterized by suspiciously high turnouts – especially among individuals below the age of 25 – and relatively low violence overall. Nevertheless, several suicide attacks and bombings resulted in the closing of some polling centers.

Since no candidate was able to achieve an outright majority in the first round, a runoff election had to be held between the front-runners of this round: Abdullah and Ghani. In the following subsection, we will give brief descriptions of both of these candidates, as well as a summary of how their ethnicities factored into their popularities.

2.2 Runoff Election Candidates and Ethnicity Factors

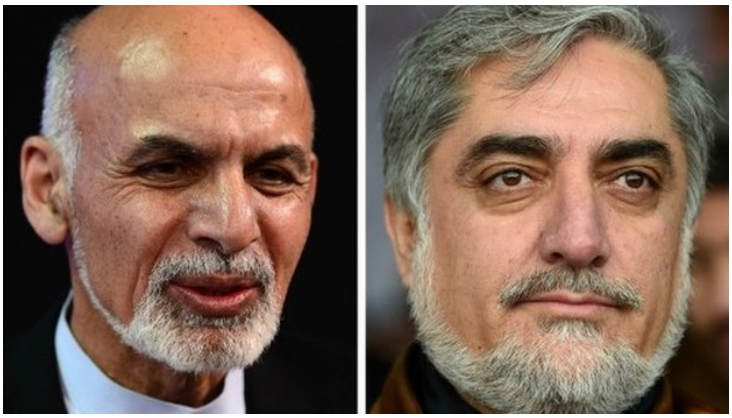


Figure 2: *Picture of Ghani (left) and Abdullah (right). Image adapted from [11].*

2.2.1 Abdullah Abdullah

Dr. Abdullah Abdullah, born in Kabul in 1960, is half-Pashtun and half-Tajik. While working as an ophthalmologist during the Soviet War in Afghanistan, he advised Mujahideen commander Ahmad Shah Massoud and provided medical aid to civilians. Abdullah went on to hold spokesman roles during the Taliban regime, and was eventually appointed Minister of Foreign Affairs under Karzai in 2001. After resigning from that position in 2005 and dropping out of the political scene for a few years, he reappeared in 2009 as a candidate in the presidential election, competing against the incumbent Karzai.

Following the 2009 election's first round of voting, it was initially reported that Karzai achieved a majority of 54.6% of the vote. However, in the wake of fraud allegations, the UN-dominated Afghan Independent Electoral Complaints Commission (ECC) decided to audit the election's results. Following the audit, Karzai was found to have won 49.67% of the vote, while Abdullah won 30.59%. Since Karzai fell short of a majority by 0.33%, he and Abdullah negotiated and decided to hold a head-to-head run-off election on November 7. However, on November 1, Abdullah withdrew his candidacy, since he did not trust Karzai's government's ability to hold a free and fair second round.

Following the 2009 election, Abdullah created the "National Coalition of Afghanistan" – a grassroots party that campaigned against the political ideologies of Karzai.

2.2.2 Ashraf Ghani

Dr. Ashraf Ghani, born in Logar in 1949, is Pashtun. Ghani holds a PhD in Cultural Anthropology from Columbia University, specializing in social transformation and state-building. Following the destruction of the Twin Towers, he helped the UN prepare the Bonn Agreement, which re-created the State of Afghanistan. This agreement called for a democratic political system; a Supreme Court; and a judicial system in accordance with Islamic principles, Afghan traditions, and international standards. Following the establishment of the Bonn Agreement, Ghani went on to become the chief advisor to Karzai, the Chancellor of Kabul University, and the Finance Minister of Afghanistan.

Ghani also campaigned in the 2009 presidential election; with only 3% of the total vote, he placed in fourth out of thirty-eight candidates. During this campaign, he promoted the development of a representative government built on public feedback.

2.2.3 The Ethnicity Factor

One aspect worth noting is the relationship between Ghani and the Pashtun provinces of Paktya, Khost, and Paktika. Ghani, who hails from the neighboring Logar province, is connected to these regions by his Pashtun background. Due to the comparatively low populations of these provinces, individuals there are poorly represented. So, to the citizens of those regions, the Pashtun Ghani's presidential bid promises increased political representation.

The half-Tajik Abdullah, on the other hand, was raised in Panjshir. To the people of Paktya, Khost, and Paktika, he represents the threat of a Panjshiri-dominated government. Since the Panjshiris overran the security ministries in the post-Taliban era, and did a poor job at controlling crime in and around Paktya, this threat carries a significant amount of weight in the eyes of the Paktiawals [9]. By placing Ghani in power, the perceived result would be more political representation and more security in Paktya, Khost, and Paktika. Consequently, these three provinces overwhelmingly supported Ghani over the Tajik Abdullah.

These provinces were also notorious for extensive fraud in the 2009 presidential election, through fraudulent proxy female ballots [9]. It is therefore not surprising that – in the run-up to the runoff election – numerous mullahs in these regions traveled between villages and preached a religious obligation to vote for Ghani. In fact, several villages were found to impose fines on – and make violent threats at – those who openly voted for Abdullah. Similar reports describe village officials making agreements with the Taliban about voting unanimously for Ghani, in exchange for a peaceful election day.

2.3 Runoff Election and Aftermath

The runoff election between Ghani and Abdullah was held on June 14, 2014. Due to the many accounts of suspected fraud in the preliminary round, observer deployment increased across all provinces.

Of the 20 million registered voters in Afghanistan, only 8 million votes were counted, resulting in an "overall voter turnout" of 38.9%. However, this registered voter count is inflated by fraudulent proxy over-registration [13] and diaspora voters. It makes sense to instead look at the percentage of the voting-age population (VAP) that voted, and take that as the "official" turnout figure. According to [14], this VAP turnout percentage was 50.03% – which is significantly different from the "overall voter turnout" mentioned earlier.

The preliminary results of this election suggested that Ghani had won with 56.44% of the vote. However, several eyewitness accounts described rampant amounts of fraudulent voting during this runoff round. For instance, in the province of Khost, a pro-Abdullah team led by Manawar Shah reportedly saw 500 Ghani supporters (including election officials) cast 10,531 votes via multiple ballots [12]. Shah's team alleges that they could not document this fraud since they had been beaten and prevented from using their cameras or phones; all they could do was watch. In other reports, some election officials reputedly attacked Abdullah's supporters and broke their recording equipment. All of these narratives of pro-Ghani fraud are corroborated by several intercepted phone calls, wherein Ghani campaign workers directed officials to stuff ballot boxes.

Based on these eyewitness accounts, Abdullah accused Ghani supporters, government officials (including Karzai), and the IEC itself of committing large-scale fraud. After an extensive, EU-led, three-month audit, the IEC declared Ghani the new president of Afghanistan on September 21, 2014, with roughly a one million vote lead above Abdullah (i.e. 54% to 44%).

As part of this announcement, Abdullah and Ghani also signed a power-sharing agreement. With this agreement, Abdullah acquired the political title of "Chief Executive Officer of Afghanistan" – a new position that is briefly described in Section 1.2.

As agreed in Abdullah and Ghani's power-sharing agreement, the IEC decided that it would not release the exact, "fraud-free" vote counts from the runoff election. However, auditing results suggest more than two million votes were potentially fraudulent, even with 1.7% of all votes previously disqualified as invalid [15].

3 Objectives for Fraud Analysis

Over the course of this paper, we use a variety of techniques to analyze data from the 2014 Afghan presidential election and search for fraudulent patterns. We address two main questions in our analysis:

1. Do the runoff election's results show signs of fraud, and – if so – in favor of which candidate(s)?
2. If there was fraud, where did it occur?

In Section 5, we analyze the election results and look at the data at polling-station, district, and province levels of aggregation. This analysis involves various fraud indicators, as summarized at the beginning of this section, and these indicators are used to narrow in on both the mechanisms and perpetrators of fraud. In Section 6, we inspect regions with abnormally high turnouts and look at skewed vote-share distributions. This investigation helps us answer the second question of where exactly the fraud occurred.

4 Data

Afghanistan is split into 34 provinces, each of which contains between 5 and 28 districts. In total, there are ~ 400 officially recognized districts in the country, although this number is constantly changing as provinces reorganize themselves. During the 2014 election, each district set up between ten and several hundred polling centers, in environments as diverse as mosques, schools, and private houses [16]. Each polling center could have contained up to 12 polling stations, with a bare minimum of 2 polling stations required (one for men and one for women). Due to cultural norms, these polling stations were segregated by gender; the IEC had designated 8,958 of 21,663 polling stations (i.e. 41.4%) as female-only in this election [17].

4.1 Election Data

We analyzed CSV data on both rounds' vote counts (at the polling-station level) and observer deployment numbers (at the district level) from the website of the Afghanistan Election Data project [18]. This website's runoff election results are just the preliminary numbers that were released soon after the election. Note that these are not the "clean" vote counts from the EU's auditing process. As discussed in Section 2.3, the power-sharing agreement between Ghani and Abdullah required the IEC to not disclose the audited voting numbers.

Unfortunately, the election data for both elections had many issues. First, several entries in the polling-station-level runoff data had districts and provinces marked as "NA", even if those stations had recorded vote counts. By comparing the polling center IDs of these entries with the first-round data, we determined that many of these "NA" entries corresponded to "temporary" districts (due to reorganization). For example, six of these "NA" entries in the runoff data had polling center IDs of 2717011. By comparing polling center IDs against the first-round data, we noticed that these entries were from the Dand district in Kandahar Province, which was marked as "temporary" in our population data. Because these data were marked as "not applicable," we were not able to use them in our analysis.

Furthermore, there were several issues with inconsistent (and sometimes wildly-varying) province/district spellings. These issues had to be resolved by comparing polling center IDs and deciding on a single spelling of each province and district. Throughout our analysis, we try to remain consistent in our use of province and district names.

Lastly, we noticed that the Gizab District was marked as being in Daykundi Province, even though it was actually re-annexed into Urozgan Province in 2006 [19].

4.2 Population Data

We used XLSX data on each district's settled population (recorded to the nearest thousand people) from the 2013-14 report published by Afghanistan's Central Statistics Organization (CSO) [20]. Since voter registry data is not available on a district-level resolution, we instead use these population data to calculate the turnout. Specifically, we estimate the eligible number of voters in a given district by taking that district's population and multiplying it by 0.5093 – the fraction of the Afghan population that is estimated to be of voting age [14].

It is also worth noting that Afghanistan has not completed a full census in several years. The population figures used by the CSO are projections and corrections built on the results of a 1979 census, but this census was never actually finished [13]. As a result, there are most likely several inaccuracies in the CSO's data.

Moreover, there are many districts in the 2013-14 CSO data that are marked as "temporary." These districts hadn't actually been created yet, but for some reason they had population data associated with them. In calculating the population of each province, we have ignored these temporary districts, since their election data was marked as "not applicable" (as mentioned in Section 4.1).

In spite of all these imperfections, the CSO is the only official source of Afghan population data on a district-level resolution. Thus, its statistics will have to suffice for the purposes of this analysis.

4.3 Province Number to Name Mapping

In some of our province-level histograms (see, for example, Section 5.2), we use numbers to represent provinces. This bijective mapping between province numbers and names is covered in Table 1.

Province Number	Province Name	Province Number	Province Name	Province Number	Province Name
0	Badakhshan	12	Jawzjan	24	Paktika
1	Badghis	13	Kabul	25	Paktya
2	Baghlan	14	Kandahar	26	Panjshir
3	Balkh	15	Kapisa	27	Parwan
4	Bamyan	16	Khost	28	Samangan
5	Daykundi	17	Kunar	29	Sar-E-Pul
6	Farah	18	Kunduz	30	Takhar
7	Faryab	19	Laghman	31	Urozgan
8	Ghazni	20	Logar	32	Wardak
9	Ghor	21	Nangarhar	33	Zabul
10	Helmand	22	Nimroz		
11	Herat	23	Nooristan		

Table 1: *The mapping from province numbers to province names.*

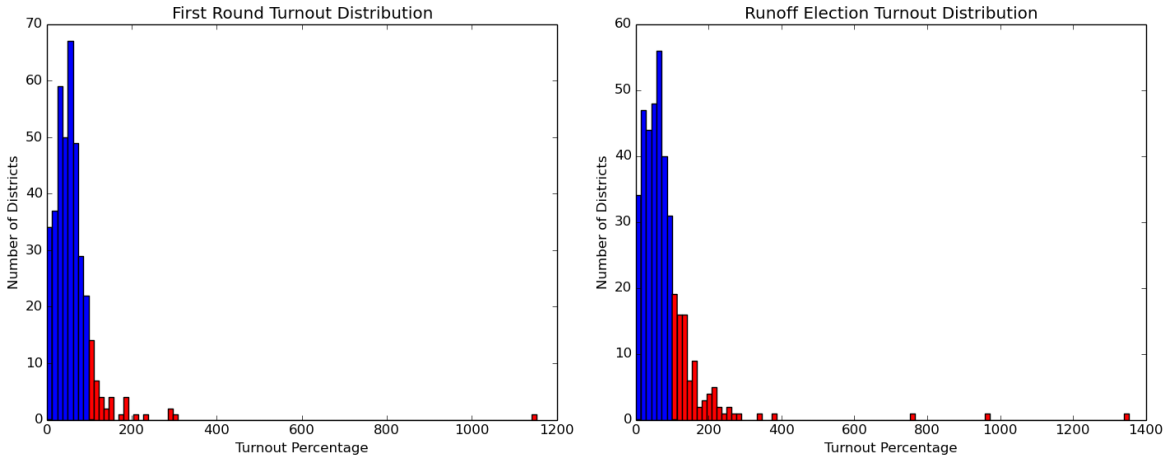
5 Country-Wide Fraud Analysis

In this section, we focus on analyzing the entire country's election data, on both province-level and district-level resolutions. To search for signs of fraud and their perpetrators, we use various indicators, such as turnout distributions (Section 5.1); a winning margin analysis (Section 5.2); electorate share vs turnout plots (Section 5.3); vote share vs turnout plots (Section 5.4); correlations between turnout and observer deployment (Section 5.5); and digit analysis (Sections 5.6 and 5.7).

5.1 Turnout Distributions

Turnout distribution histograms describe the frequencies of various turnout rates. In a free and fair runoff election, we would expect these distributions to be approximately Gaussian and centered on the election's average turnout rate of 50.03% (of the VAP) [14]. Deviations from this symmetric, Gaussian shape serve as potential indicators of fraud. For example, a bimodal turnout distribution would suggest significant ballot stuffing in some regions (barring other explanations based on demographics). Such distributions, however, tend to be rare in practice, since it would take a large amount of fraud to create a second peak in a turnout distribution. Instead, it is far more common to see significant skews in fraudulent turnout histograms [21].

To calculate the turnout in a district, we take the total number of votes in that district and divide by its voting-eligible population, which is 0.5093 times its CSO-listed population. As mentioned in Section 4, this is the only approach that we can realistically take; since there is no openly-available voter registry for us to use [22], we must assume that a constant fraction of each district's population is eligible to vote. Through this procedure, we created district-level turnout distributions for both the first round and runoff round; these are shown in Figures 3a and 3b, respectively:



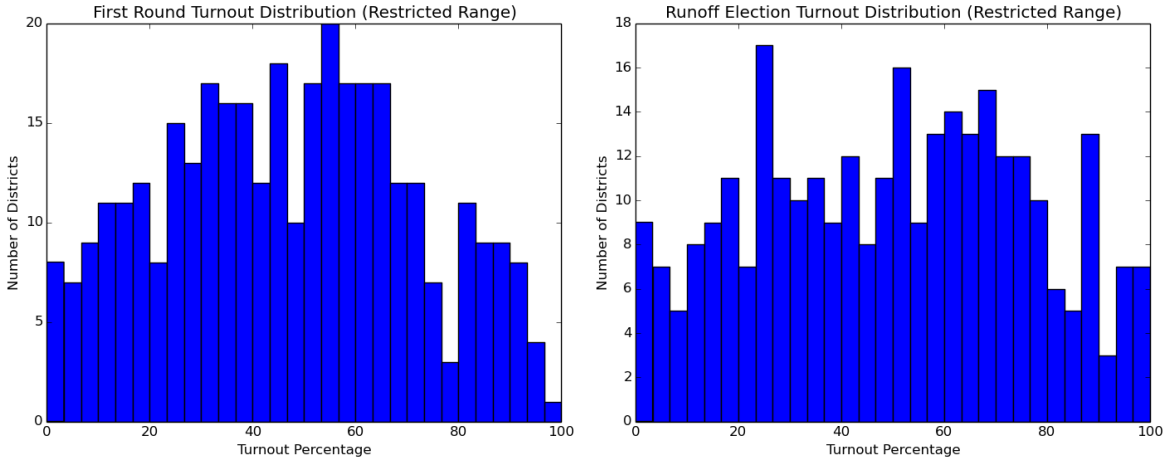
(a) *First-Round Turnout Distribution* **(b)** *Runoff Turnout Distribution*

Figure 3: *District-level turnout distributions for the first round and runoff election. Districts with* $> 100\%$ *turnout are marked in red. Note the non-Gaussian nature and skew of both histograms.*

Note that both of these turnout distributions have long tails at high turnout percentages, and there are several districts with $> 100\%$ turnout (which are marked in red in Figure 3). The sheer number of districts with $> 100\%$ turnout (in both elections) might seem suspicious at first, but we must remember that the CSO's population data is not entirely accurate since no comprehensive census of Afghanistan has been completed in several years [13]. Also, the voting-eligible fraction of 50.93% might not apply equally to all districts. Lastly, it's important to note that the VAP fraction of 50.93% does not include diaspora voters. This is problematic since, as Section 6.2 shows, many of the high-turnout regions also happen to be in provinces on the border with Pakistan. Afghan refugees or nomads might have entered (or temporarily moved to) Afghanistan to vote, even if they don't constitute the VAP.

However, even after we consider all of these caveats, the outrageously high turnout percentages in these histograms still aren't fully accounted for. This suggests that ballot stuffing *might have* occurred in various districts and provinces throughout the country – during both election rounds. In addition, based on Figure 3, we can actually see that there was a general increase in turnout between the first round and runoff election, with several more districts reporting $> 100\%$ turnout in the runoff election. Some of this discrepancy can be attributed to the tribally-organized, pro-Ghani voter mobilization efforts in Pashtun provinces (e.g. Paktya, Khost, and Paktika) before the runoff election [9]. Yet it is unlikely that organic mobilization (by itself) produced such impossibly large turnout percentages. Instead, increased ballot stuffing might *partly* be responsible for the runoff election's turnout distribution, especially given the higher stakes of this second round.

As a final point of comparison between the two rounds' turnout distributions, we've zoomed in on the [0, 100] percent domain of the histograms in Figure 3; the resulting plots are shown in Figure 4:



(a) *First-Round (Restricted) Turnout Distribution* **(b)** *Runoff (Restricted) Turnout Distribution*

Figure 4: *Restricted district-level turnout distributions for the first round and runoff election. Note that only districts with* $< 100\%$ *turnout are included in these histograms.*

Clearly, both of these histograms exhibit signs of non-normality due to their asymmetry. Yet perhaps more importantly, the runoff election's histogram is significantly less Gaussian than the first-round histogram, since the former lacks a clear peak near the 50% mark (the VAP turnout for the second round [14]). Of course, this non-unimodal shape could result from legitimate voter rallies in tribal areas (as mentioned earlier), but it's also possible that fraud is to blame for the non-Gaussianity.

Overall, the turnout distributions for this election are not clearly indicative of fraud, yet they do raise many questions about the runoff election's legitimacy. We will have to use more sophisticated fraud indicators if we want to carefully isolate and analyze fingerprints of fraud.

5.2 Winning Margin Analysis (WMA)

A winning margin analysis (WMA) involves comparing the winning candidate's margin of victory in a region (i.e. the percentage by which he won that region) against the turnout percentage in that region. If provinces with large winning margins (for Ghani) consistently have higher-than-average turnouts, then this would imply that Ghani is asymmetrically favored by high turnouts across the country. Specifically, if this correlation between winning margin and turnout is strongly linear and positive, this would suggest that the extra turnout was funneled directly into stuffed ballots for Ghani, which is indicative of large-scale fraud. On the other hand, if there is no net correlation between Ghani's winning margin and turnout, all we can conclude is that there isn't any *nationwide* fraud being committed by Ghani (there might still be district-level fraud, and/or some amount of fraud being committed by Abdullah).

In Figure 5, we have plotted Ghani's runoff winning margin in each province and the turnout in that province minus 50% (the approximate average VAP turnout) in a combined bar graph. Here, the winning margin is calculated by subtracting Abdullah's percentage of the vote in that province from Ghani's percentage of the vote.

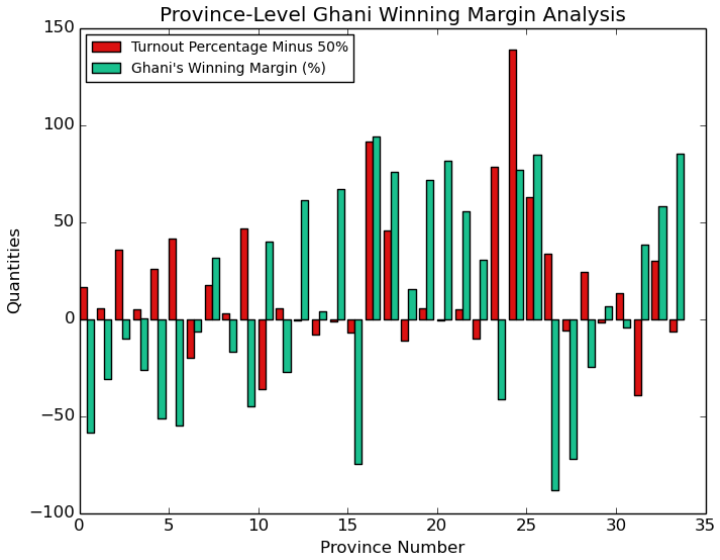


Figure 5: *The turnout percentage (minus 50%) and Ghani's winning margin (as a percentage) for each province. Note that there is no general, positive correlation between these two quantities. N.B. the mapping from province numbers to province names is given in Table 1, and province numbers start at zero.*

Clearly, there is no general, positive correlation between Ghani's winning margin and the turnout in each province. This would suggest that massive, country-wide fraud wasn't committed by Ghani's side (and *only* Ghani's side) during the runoff election. At the same time, this winning margin analysis does not rule out the possibility of more localized fraud. There are a number of suspicious provinces where Ghani had both a high turnout (of over 100%) and a high winning margin, such as Provinces 16, 24, and 25. According to Table 1, these correspond to Khost, Paktika, and Paktya (respectively). We will want to keep a close eye on these provinces in Section 6, since it is likely that Ghani engaged in ballot stuffing there.

5.3 Share of the Eligible Electorate vs. Turnout (V/E vs T)

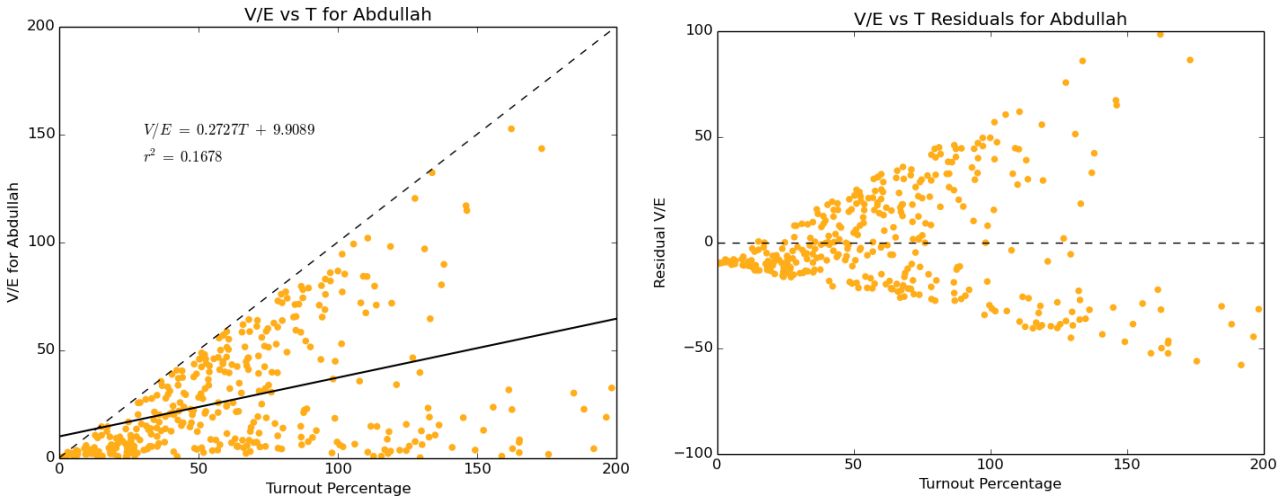
In looking for fraud, it is useful to see how the turnout in a district affects the share of the eligible electorate that voted for a candidate there. This analysis can be done by creating V/E vs T plots for the runoff election, where:

- V is the number of votes for a given candidate in a certain district.

- E is the number of voting-eligible citizens in that district. We assume that this is equal to $0.5093P$, where P is the district's population.
- T is the turnout percentage for that district. As before, this is calculated by taking the total number of votes and dividing by $0.5093P$.

In a given district, we would expect V/E to increase (approximately) linearly with T. After all, if 40% of voters support Candidate X in a given district, then an increase in turnout by Y votes should result in (around) $0.4Y$ more votes for Candidate X. A typical V/E vs T analysis also assumes that there is no significant correlation between voting preferences and geographic location. Under this assumption, we can combine data from various districts into a single plot of V/E vs T. For a candidate in a free and fair election, then, a fitted line on such a plot should have a vertical intercept of 0 (since no turnout means no votes) and a slope equal to that candidate's country-wide vote share (0.5644 for Ghani, and 0.4356 for Abdullah).

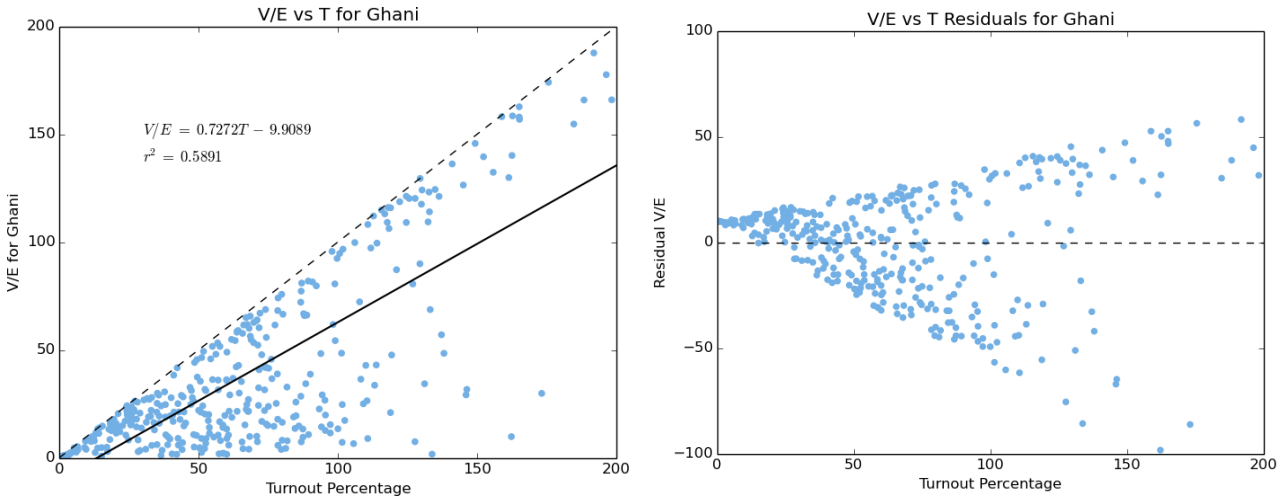
In the following two figures, we've plotted V/E vs T for both Abdullah and Ghani, and have included fit lines (in thick black) and residual plots in both cases. Note that these plots and fits ignore districts with turnouts greater than 200%, since those few very-high-turnout districts would be outliers in our linear fit.



(a) *Abdullah's V/E vs T Plot and Linear Fit*

(b) *Residuals*

Figure 6: *V/E vs T analysis for Abdullah. We've included a plot of the data with a linear fit (in thick black), as well as a residual plot. The diagonal line $V/E = T$ has also been plotted for reference. While the best-fit line does have a positive slope, the slope doesn't match Abdullah's vote share (0.4356).*



(a) *Ghani's V/E vs T Plot and Linear Fit*

(b) *Residuals*

Figure 7: *V/E vs T fit plot (with the best-fit line in thick black) and residuals for Ghani. The diagonal line $V/E = T$ has also been plotted for reference. While the best-fit line does have a positive slope, this slope doesn't match Ghani's vote share (0.5644).*

Clearly, while both of these plots have intercepts close to zero, they do not have the expected slopes. Statistically, this is not a big surprise; the residual plots show that our data (for both candidates) is broadly separated into two sections: districts that lie *very* close to the line $V/E = T$ (i.e. those where the candidate got almost all of the district's votes), and districts that lie close to $V/E = 0$ (where the candidate got almost none of the district's votes). Thus, it is no real surprise that our V/E vs T plot fails to match the theory so exactly.

Putting aside the statistics of least-squares fits for the moment, one possible non-fraudulent reason for this strange V/E vs T behavior is heterogeneity in the data. Afghanistan is filled with ethnicity-driven politics, as evidenced by the Pashtuns' fears of Panjshiri political dominance [9]. As a result, it doesn't make sense to assume a lack of correlation between voting habits and location. Furthermore, we might not expect the marginal increase in a candidate's vote count to vary so linearly with turnout – especially since the first 20% of voters may not necessarily vote in the same way as the last 20% would.

Nonetheless, these concerns do *not* make our V/E vs T plots useless as fraud indicators. For one thing, it's highly suspicious that there are several districts where V/E increases 1:1 with turnout, for both candidates. One could claim that some districts just unanimously support a given candidate, but this would not account for the sheer number of data points that lie so close to $V/E = T$. Thus, if we ignore the possibility of very strong, ethnically-driven preferences, Figures 6 and 7 *suggest* that both Ghani and Abdullah *might have* been involved in ballot stuffing.

The more clear-cut culprit, based solely on these plots, is Ashraf Ghani. After all, Ghani has several districts where he won almost 100% of the vote share (i.e. $V/E \approx T$) and had over 100% turnout. This suggests that, in those districts, pro-Ghani supporters (possibly organized by the district's mullahs or election officials) engaged in rampant amounts of ballot stuffing. As mentioned in Section 6, most of these high-turnout districts were in Pashtun provinces (e.g. Paktika, Paktya, and Khost), where Ghani already had a significant amount of tribal support through voter mobilization [9]. Thus, it is not unreasonable to claim that pro-Ghani mullahs or officials were able to orchestrate and get away with such high amounts of fraud in those districts.

One could argue that the $> 100\%$ turnout aspect is just an artifact of our turnout calculations' inaccuracy, and/or the inaccuracy of the CSO's population data. Yet even if we treat our turnout calculations as *approximately* correct, it's clear that Ghani had high vote shares in more high-turnout areas than Abdullah did. This fact, by itself, is enough to raise a few flags.

In conclusion, then, this V/E vs T analysis suggests that Ghani did engage in ballot stuffing in many districts; this is consistent with the eyewitness anecdotes mentioned in Section 2. Abdullah, on the other hand, might have engaged in some amount of ballot stuffing, based on the number of districts where he has very high shares of the vote; however, it is not entirely obvious that Abdullah is as culpable as Ghani in this regard.

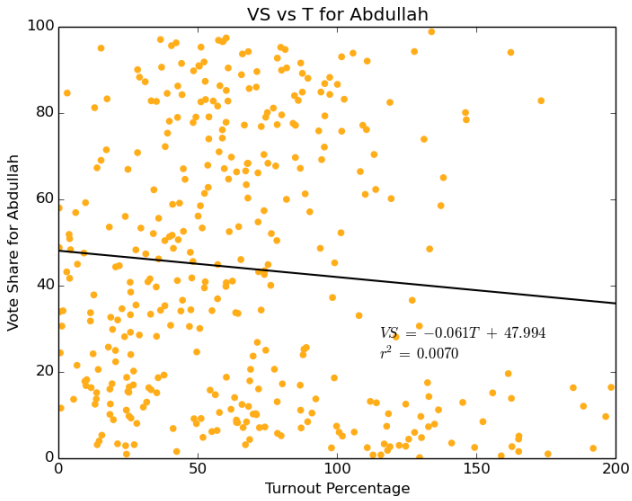
So far, our analysis has focused on indicators related to ballot stuffing. In the following section, we will consider the distribution of each candidate's vote shares as a function of turnout, in order to investigate vote stealing.

5.4 Vote Share vs. Turnout (VS vs T)

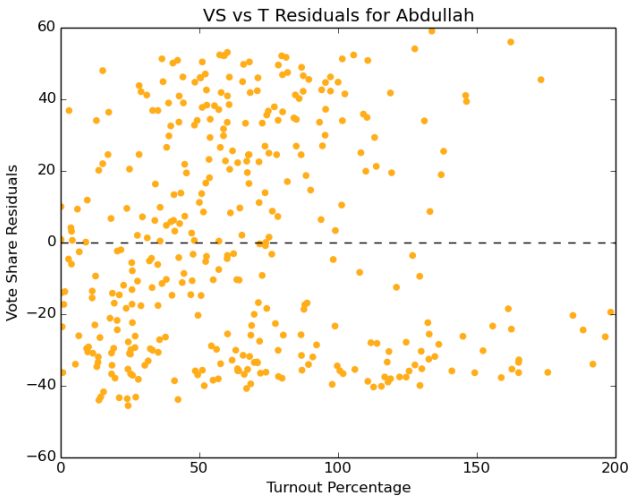
A candidate's vote share (VS) in a given district is defined as the fraction of votes that he won there. If we were looking at a non-fraudulent election in a politically homogeneous country, we would expect there to be no net correlation between VS and T (turnout). After all, the fraction of votes that go towards a candidate in a given district shouldn't depend very strongly on how many people show up to vote. If we extend this reasoning to the entire country (by the homogeneity assumption), we can see that we wouldn't expect a nationwide correlation between VS and T. Any deviation from this would be indicative of fraud.

Even if we relax the homogeneity condition (so that vote shares vary significantly across the country), a strong, negative correlation between Candidate X's vote share and turnout would be highly suspicious. This means that Candidate X primarily lost votes to his opponent Candidate Y in high-turnout regions. Thus, unless Candidate Y is legitimately preferred in high-turnout districts (for demographic reasons), this negative correlation implies that Candidate Y stole votes from Candidate X in order to boost his vote totals in key, high-turnout areas.

In Figures 8 and 9, we have used district-level voting and population data to create VS vs T plots for Abdullah and Ghani (respectively). A linear fit has been included in order to study correlations, as well as residual plots to analyze deviations. Both the graphs and linear fits only use data from districts where turnout is less than 200%; as mentioned in Section 5.3, very-high-turnout regions would serve as outliers in such a fit.

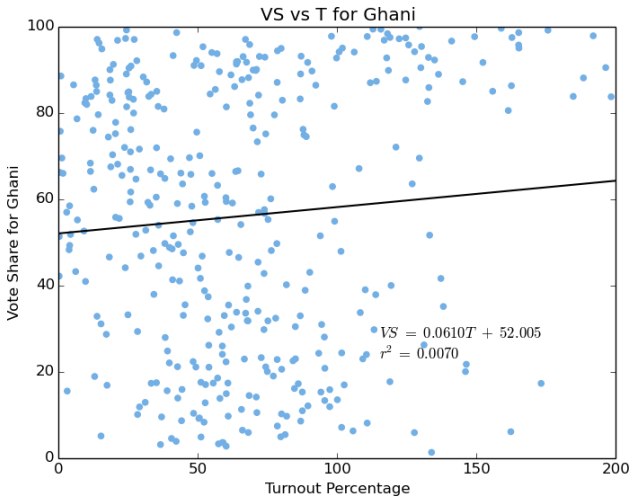


(a) *Abdullah's VS vs T Plot and Linear Fit*

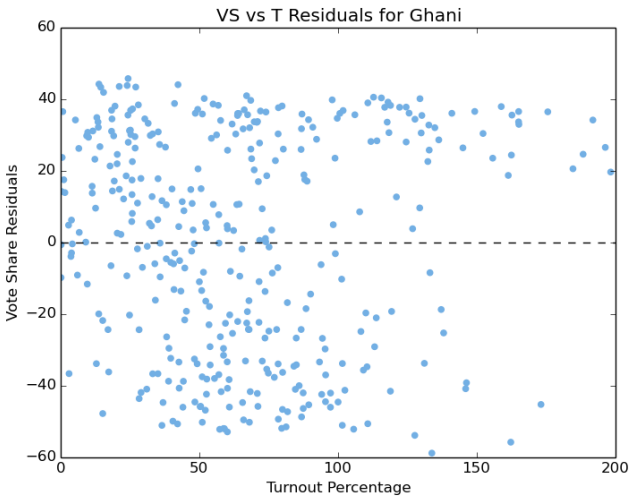


(b) *Residuals*

Figure 8: *VS vs T fit plot (with the best-fit line in thick black) and residuals for Abdullah.*



(a) *Ghani's VS vs T Plot and Linear Fit*



(b) *Residuals*

Figure 9: *VS vs T fit plot (with the best-fit line in thick black) and residuals for Ghani.*

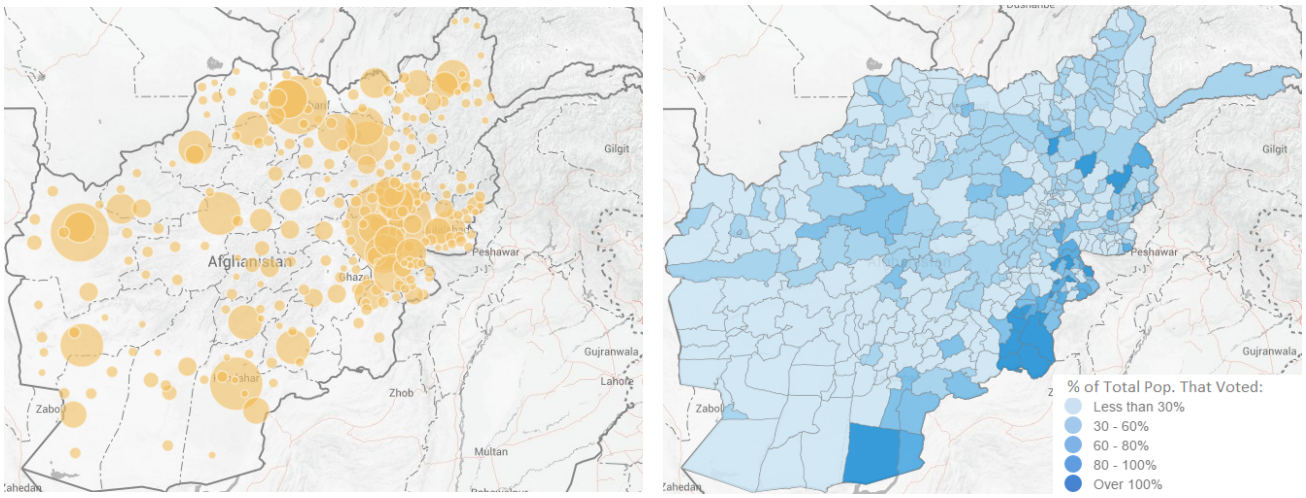
Based on the r^2 values of both linear fits, we can see that there is no strong net correlation between each candidate's vote share and turnout. At the very least, this tells us that there wasn't a rampant amount of nationwide vote stealing by either candidate; if there were, we would expect a much clearer, negative correlation in the losing candidate's plot.

Nonetheless, there are a few interesting points to note about these two figures. First, we can see that the residuals are evenly distributed about the fit line in the $< 100\%$ turnout domain. This tells us that there was no net correlation between a candidate's vote share and turnout in this (relatively) "low-turnout" regime, which is what we'd expect in a free and fair election. However, in the $> 100\%$ turnout (i.e. "high-turnout") domain, we can see that Ghani won more districts than Abdullah did. This once again shows that Ghani was the outright winner in most high-turnout districts. On the other hand, Abdullah only won in a few of these high-turnout districts – and often by a much smaller margin than 100% of the vote share.

Even if our turnout calculations are only approximately accurate, the clear divide between the $> 100\%$ turnout regions and the $< 100\%$ turnout regions is highly unusual. Specifically, since the suspicious regions in our VS vs T residual plots are almost exclusively those with $> 100\%$ turnout, we can infer that pro-Ghani supporters perpetuated fraud via ballot stuffing, not vote stealing. While Abdullah's engagement in ballot stuffing remains less clear, the data does suggest that he received large fractions of the vote share in a couple of high-turnout regions.

5.5 Turnout and Observer Deployment Correlations

An increased number of election observers were sent to moderate the runoff election. However, as shown in Figure 10a, there were several gaps in observer deployment throughout the country. Perhaps more interestingly, as shown in Figure 10b, the districts with relatively low observer deployment also had impossibly high turnout levels – sometimes more than 100% of their CSO-listed populations. This suggests that fraud was committed in areas with relatively low observer deployments – presumably because it was easier for the fraudsters to "get away with it."



(a) *Observer Deployment Map* **(b)** *Turnout (as a Fraction of Total Population) Map*

Figure 10: *Observer deployment and turnout maps for the runoff election, modified from [23]. In the observer deployment map, larger yellow circles correspond to higher observer deployments. In the turnout map, darker shades of blue correspond to higher turnouts, with the darkest shade corresponding to votes by* $> 100\%$ *of the total population.*

While this seems like a plausible explanation for the prevalence of fraud along the Pakistani border, it's possible that high turnout and low observer deployment just happen to coincide in the runoff election data. In other words, it's possible that this isn't indicative of an underlying trend. Moreover, we should really be comparing the observer deployment density, since we'd expect fewer observers in less-populous regions.

Fortunately, we can resolve some of these ambiguities by considering the change in observer deployment density between the first-round and runoff elections; analyzing this change would help remove other confounding factors and focus on the correlation at hand. Specifically, we'll look at the relative observer deployment density percent change, or RODDPC. To compute this quantity for each province, we first take the number of observers in a given province, divide by that province's population (to get a dimensionless, properly-scaled "density"), and measure the percent change in this quantity between the first and second rounds. We then subtract off the median observer deployment density percent change (across all 34 provinces) to get the RODDPC (hence the "relative" descriptor). By subtracting off the median, we can look at which provinces gained relatively more observers than others between the two rounds; after all, we already know that almost all provinces had more observers for the second election.

In the following figure, we've plotted the RODDPC and the percent change in turnout between the two election rounds, for each province:

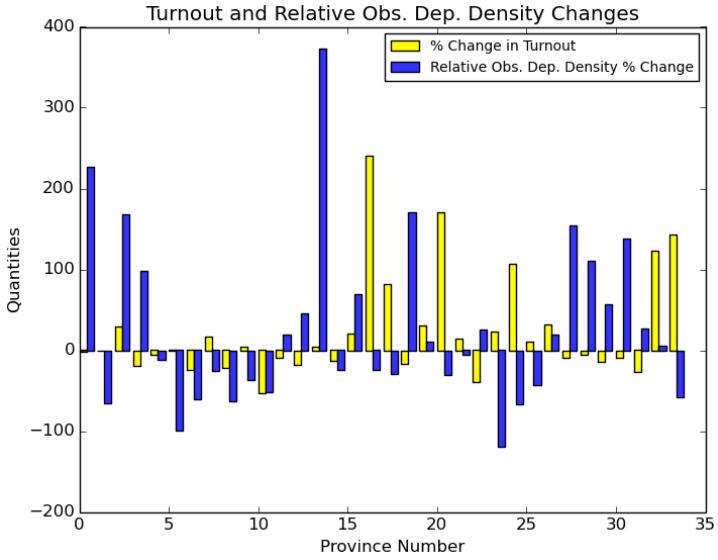


Figure 11: *The percent change in turnout and RODDPC (defined above) for each province, measured between the first-round and runoff elections. In many provinces, these two quantities change in opposite directions, especially in those where turnout went up between the two rounds. N.B. the mapping from province numbers to province names is given in Table 1, and province numbers start at zero.*

For the most part, it appears as if there is a negative (albeit nonlinear) correlation between RODDPC and turnout, since the bars in Figure 11 tend to move in opposite directions for each province. Furthermore, provinces with large increases in turnout almost always coincide with provinces where the observer deployment density didn't increase as much (i.e. where the RODDPC was negative), with the only real exception being Province 32 (Wardak, which had a very small, positive RODDPC). Based on these trends, we can propose a reasonable mechanism for fraud in the Afghan elections. Namely, the perpetrators of ballot stuffing might have known the observer deployment levels of the second round well in advance, and so they may have purposely targeted the regions that were (relatively) less well-observed compared to the previous round. This targeting certainly wasn't perfect, since there are some regions in Figure 11 (e.g. Province 5 – Daykundi) with somewhat large dips in RODDPC but only minuscule changes in turnout. Yet these imperfections could be attributed to the "prior distribution" of fraud before the runoff; fraudsters might have been more concentrated in certain areas (e.g. Paktika, as mentioned in Section 6), and hence wouldn't have been able to travel across the entire country en masse to stuff ballots.

Of course, it's also possible that Figure 11 merely indicates correlation instead of outright causation. One could argue that there are additional factors at work, such as increased pro-Ghani mobilization efforts that coincidentally took place in provinces with (relatively) few observers. However, the aforementioned mechanism for fraud does match up with the anecdotal reports detailed in Section 2.3, where eyewitnesses describe pro-Ghani election officials repeatedly casting votes in provinces like Khost. Note that Khost is Province 16, which has the tallest yellow bar in Figure 11, as well as a RODDPC of -23.88%. Thus, our proposed mechanism for fraud does have some credibility.

5.6 First-Digit (Benford's Law) Analysis

So far in our analysis, we have primarily looked at indicators that highlight vote stuffing and stealing. Digit analysis, on the other hand, is meant to help us catch a third, common type of electoral fraud: falsified vote totals.

One particular tool we can use as a fraud indicator is Benford's Law. This phenomenological law states that the first digits of non-systematically-generated numbers tend to follow a logarithmic pattern. Specifically, if we have a set S of base-10 numbers, that set is said to follow Benford's law if the leading digit d occurs with probability:

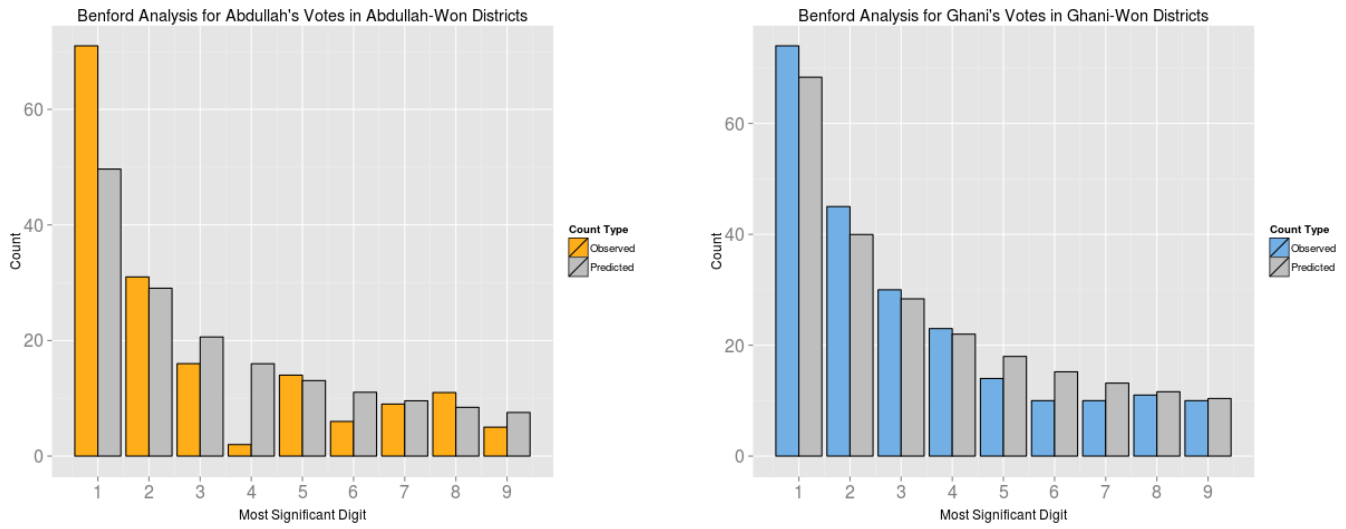
$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right) \tag{1}$$

It turns out that many types of data satisfy Benford's Law, including mathematical constants found in nature (regardless of the units in which they are expressed). In fact, as long as the data follows no other systematic process, covers several orders of magnitude, and follows a probability distribution that varies smoothly and slowly on log scales, it is likely that it will follow Benford's Law.

Since many of these conditions are not very strict, we might expect the number of votes for a certain candidate – at a given aggregation level (e.g. the district level) – to follow Benford's Law in a free and fair election. Any deviations from this expected distribution would be suggestive of falsified vote totals. However, this isn't always true; as explained in [24], while Benford's Law often *does* hold for fraud-free election data, it isn't always a reliable indicator of fraud.

In order to actually use Benford's Law, we need access to vote count data that spans many orders of magnitude. Because of this requirement, we actually cannot use polling-station level data since those vote totals only ranged from ~ 10 votes to 600 votes (where the latter is an absolute cap on how many votes a polling station can accept). As a result, we must instead use the district-level vote counts for each candidate to conduct our Benford's Law analysis.

In the following figure, we've plotted the first-digit distribution for Abdullah's district-level vote totals in the districts where he won, and have done the same thing for Ghani's winning districts. We only focus on the districts that were won by each candidate in their respective plots, since election officials are more likely to have committed fraud in the areas controlled by their favorite candidates. The colored bars represent observed values for each candidate, while the gray bars represent the expected values (according to Benford's Law).



(a) *Abdullah's District-Level First-Digit Distribution in Districts He Won*

(b) *Ghani's District-Level First-Digit Distribution in Districts He Won*

Figure 12: *Benford's Law first-digit analysis for both candidates' vote counts, using district-level count data for the districts where each candidate won. Expected values (according to Benford's Law) are shown in gray.*

In both of these plots, we can see that the frequencies do generally tend to decrease as the most-significant digit increases. However, it appears as if Abdullah's vote totals (in the districts where he won) differ more significantly from Benford's Law, while Ghani's vote totals (in the districts where he won) fit Benford's Law fairly closely. To be more rigorous about this, we'll conduct a Pearson's chi-squared goodness-of-fit test on both distributions, using eight degrees of freedom (since there are nine digits and Benford's Law doesn't involve any additional free parameters). Our null hypotheses for both tests are of the form:

$$H_0 : \text{The first-digit distribution for Candidate X does not deviate significantly from Benford's Law.}$$

Where $X \in \{\text{Abdullah}, \text{Ghani}\}$. When we perform this chi-squared test, we get a p-value of 0.02 for Abdullah's distribution, and a p-value of 0.96 for Ghani's distribution. This means that, at the 5% significance level, we fail to reject the null hypothesis H_0 for Ghani's first-digit distribution. However, we do reject H_0 (at the 5% significance level) for Abdullah's first-digit distribution.

Naively, these chi-squared tests seem to indicate that Abdullah forged his vote totals and Ghani might not have. However, there are several issues with our Benford's Law analysis. First, as mentioned in [24], fraud can often move data in the direction of satisfying Benford's law, at least in the sense of skewing the average first digit in a vote-total distribution. Also, our plots in Figure 12 only look at data on a district-level resolution, so we do not have many data points to work with (especially since Abdullah won much fewer districts than Ghani did). Perhaps more importantly, it's much more likely that vote falsification occurred at individual polling stations and polling centers, and not at the district level. Thus, our district-level analysis might have smeared out large, polling-station-level deviations from Benford's Law, or even created spurious deviations (e.g. the low frequency of 4s in Figure 12a) that don't really mean anything.

Unfortunately, as mentioned earlier, we cannot use polling-station-level data in our Benford's Law analysis since it didn't span enough orders of magnitude and was capped at 600 votes. However, other methods of digit analysis do not have such strict requirements. In the following subsection, we will use last-digit analysis to see if we can conclusively observe vote falsification on a polling-station level.

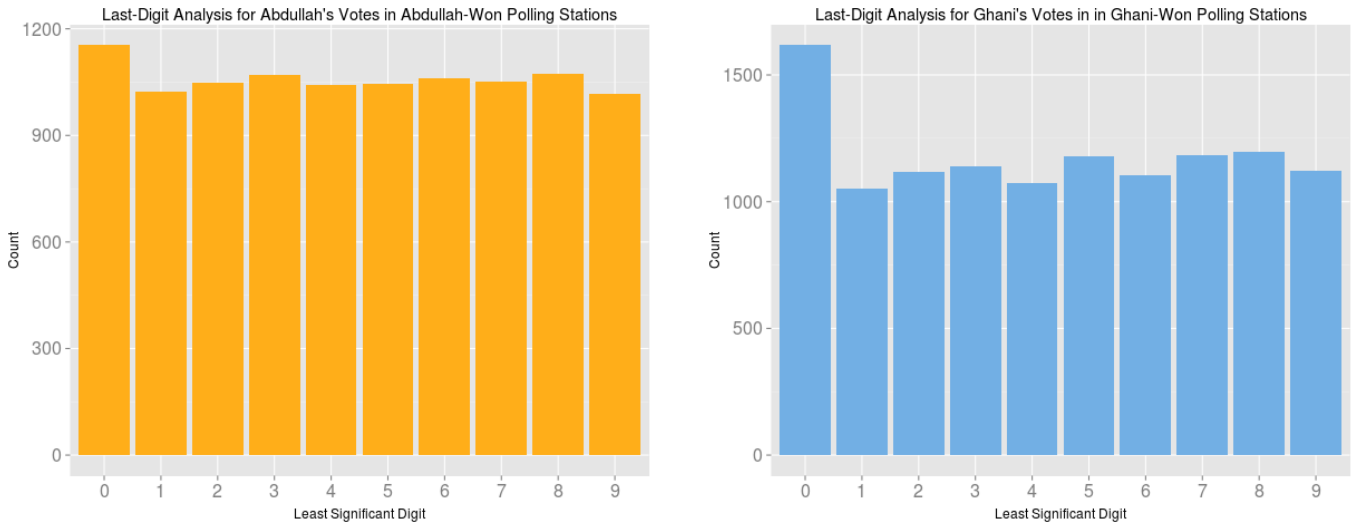
5.7 Last-Digit Analysis

In a free and fair election with accurate vote reporting, we would expect a uniform distribution of the last digits of vote totals, due to randomness and the lack of preference for any particular digit. If, on the other hand, election officials rounded digits and otherwise tampered with the data (without taking care to maintain randomness), we'd expect to see deviations from this uniform distribution. In particular, it is common for fraudulent vote totals to end with zero, just because this produces round vote totals.

Unlike Benford's law, this "last-digit law" does not require the data we use to cover several orders of magnitude. As a result, we can just use the vote counts for each candidate at the polling-station level, instead of using coarser-resolution district data. This will be more useful in helping us isolate fraud at its source: the polling stations' counts.

However, we have to be careful since polling stations in the Afghan election could only record between 0 and 600 votes each. Moreover, for the polling stations with < 10 votes for Candidate X, the last digit of Candidate X's vote count would just be the first digit, and hence would follow Benford's Law. Thus, in carrying out this last-digit analysis, we can only consider polling stations where $9 <$ (votes for Candidate X) < 600. The upper cap helps us avoid looking at polling stations that achieved their maximum vote totals, and hence had to have last digits of zero. In each candidate's last-digit distribution plots, we will also only focus on the polling stations where they won (i.e. had $> 50\%$ of the votes cast), since officials are more likely to have committed fraud in the strongholds controlled by their preferred candidates. Polling stations with tied vote counts are ignored.

In Figure 13, we include the polling-station-level last-digit distributions for Ghani and Abdullah's vote totals, after applying the aforementioned restrictions:



(a) *Abdullah's Polling-Station-Level Last-Digit Distribution in Stations He Won*

(b) *Ghani's Polling-Station-Level Last-Digit Distribution in Stations He Won*

Figure 13: *Last-digit analysis for both candidates, using polling-station-level vote data for the polling stations where each candidate won. In a fraud-free election, we would expect uniform distributions in both of these plots. N.B. in each plot, we have excluded polling stations where the respective candidate received* 600 *votes or* < 10 *votes.*

Both of these plots are highly suspicious since the digit zero appears with a particularly high frequency, even after the filtering we performed. However, Ghani's last-digit distribution is much more suspicious than Abdullah's is. In order to more rigorously quantify these plots' deviations from the expected uniform distribution, we'll conduct a Pearson's chi-squared goodness-of-fit test with nine degrees of freedom. Our null hypotheses in this test are of the form:

$$H_0: \text{ The last-digit distribution for Candidate X is uniform.}$$

Where, as before, $X \in \{\text{Abdullah}, \text{Ghani}\}$. When we perform this chi-squared test, we get a p-value of 0.73 for Abdullah's distribution, and a p-value of $< 10^{-14}$ for Ghani's distribution. This means that we fail to reject H_0 for Abdullah's distribution at the 5% significance level. However, we can reject H_0 for Ghani's distribution with a *very* high degree of statistical significance.

Thus, based on this last-digit analysis, we can deduce that election officials forged vote counts in many of the polling stations where Ghani won. More specifically, since only the digit zero appears with high frequency in Figure 13b (while the remaining digits are approximately uniformly distributed), we can conclude that these election officials rounded Ghani's vote counts to the nearest tens place, in several polling stations.

On the other hand, we cannot conclusively say much about vote count fabrication in pro-Abdullah polling stations. While the least-significant digit of zero does occur more often in these polling stations, this deviation from the expected count is not statistically significant.

6 Region-Specific Fraud Analysis

In this section, we discuss specific, problematic provinces and districts, in order to pinpoint sources of fraud. Unfortunately, we cannot produce turnout histograms for specific provinces since the CSO's population data is only available at the district level, and each province only has between 5 and 28 districts. Nonetheless, we are able to address region-specific fraud by considering the following indicators: changes in turnout between election rounds (Section 6.1); districts with very high turnouts (Section 6.2); and province-level vote-share distributions for both candidates (Section 6.3).

6.1 Changes in Turnout Between Rounds

In Figure 14, we compare the province-level changes in turnout between the preliminary and runoff rounds:

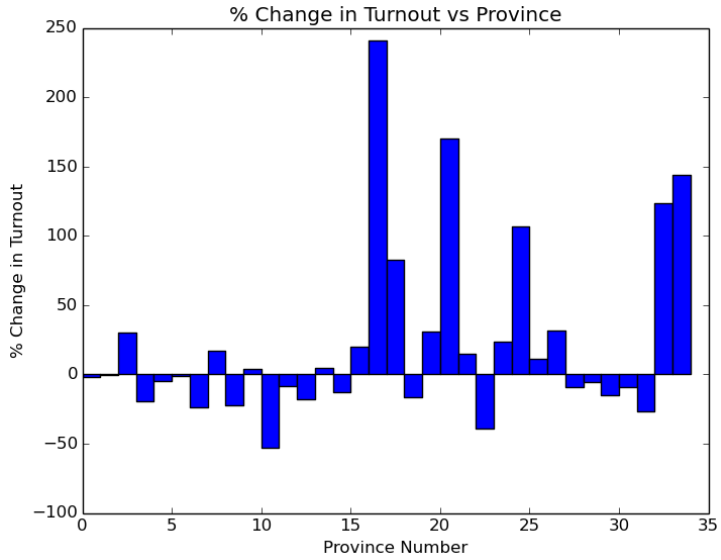


Figure 14: *The percent change in turnout between the first-round and second-round elections, by province number. N.B. the mapping from province numbers to names is given in Table 1, and province numbers start at zero.*

Based on this figure and Table 1, we can see that there are five provinces with a $> 100\%$ increase in turnout: Khost, Logar, Wardak, Zabul, and Paktika. Nader Naderi, the head of Free and Fair Election Forum of Afghanistan (FEFA), also noted this "double and in some cases three-fold increase" in votes cast in the eastern provinces [25].

These eastern provinces are the core areas of operation for the insurgent Haqqani Network, and are therefore some of the most violent areas of Afghanistan. Because of this constant threat of violence, we would not ordinarily expect high turnouts in these regions. However, according to the turnout data, Khost had about 72.0% of its CSO-listed population show up to vote, while Paktika had an overwhelming 96.2%. Since the voting-age population (VAP) forms about 50.93% of the country's population [14], this implies a turnout of $\sim 200\%$ in Paktika. Similarly, the provinces of Khost, Nooristan, and Paktya also had province-wide turnouts above 100%. Many of these provinces have histories of fraudulent voting from previous elections [26].

These rampant increases in turnout could therefore be a sign of fraud. Since the provinces of Paktika, Paktya, and Khost contributed about 970,000 votes to Ghani (94.4% of his winning margin), these possibly-fraudulent turnouts cast strong doubts on the legitimacy of Ghani's victory. However, Ghani's campaign claimed that female voter mobilization was responsible for

this surge in turnout [25]. Indeed, before the second round of the election, the tribes in these areas created an agreement in which all members of each family – including women – were encouraged to vote. This might explain the relatively large increases in turnout – especially female turnouts – during the runoff election.

However, as discussed in Section 2.2.3, this tribal agreement also mandated voting for Ghani, and specified punishments for those who supported Abdullah [9]. Moreover, many female-only polling stations, set up as a result of mobilization, experienced a shortage of female election officials. Since these polling centers were understaffed, it was difficult to keep an eye on voting irregularities. This also facilitated large amounts of fraudulent female proxy voting, in which male relatives voted repeatedly on behalf of the women in their household [25]. This fraud was not easy to prevent, since it was difficult to tell whether the female voters (i.e. the ones being represented by a proxy) were actually legitimate and eligible to vote.

Thus, the significant increase in turnout between the two elections could be partly justified by the tribes' mobilization efforts. Since many of the problematic provinces (e.g. Paktika, Paktya, and Khost) are on the border with Pakistan, it is also *possible* that nomads or diaspora voters crossed the border to take part in the elections. These voters would not show up as part of the CSO's data (as they are not settled in Afghanistan), and hence they might have artificially boosted turnouts by a bit. However, these explanations certainly do not rule out the presence of fraud, especially since the legitimacy and fairness of the mobilization efforts were questionable. In fact, as we will show in the following subsection, there were several districts with abnormally high turnouts, and it is hard to explain these values without concluding that they were fraudulent.

6.2 High-Turnout Districts

In this section, we focus on the districts with the highest runoff turnout percentages, in order to more precisely pin down the regions where fraud occurred. To determine which areas were suspicious, we aggregated the polling-station-level voting data in each district, and calculated turnouts via the method described in Section 5. By following this procedure, we found that 102 districts (out of ~ 400) had voter turnout greater than 95%.

The 20 districts with the highest voter turnout percentages are shown in Table 2:

Province	District	Turnout Percentage	Province	District	Turnout Percentage
Paktika	Wor Mamay	1352	Kunar	Asadabad	234
Paktika	Gomal	958	Nooristan	Paroon	227
Paktika	Turwo	757	Kandahar	Reg	220
Paktika	Surubi	376	Paktika	Janikhel	220
Nooristan	Wama	333	Nooristan	Barg-i Matal	218
Khost	Qalandar	288	Baghlan	Jalga	216
Paktya	Shawak	269	Khost	Jaji Maidan	212
Paktika	Zarghun Shar	257	Paktika	Waza Khwa	204
Paktya	Ahmadabad	252	Panjshir	Paryan	202
Paktika	Nika	240	Paktya	Laja Ahmad Khail	198

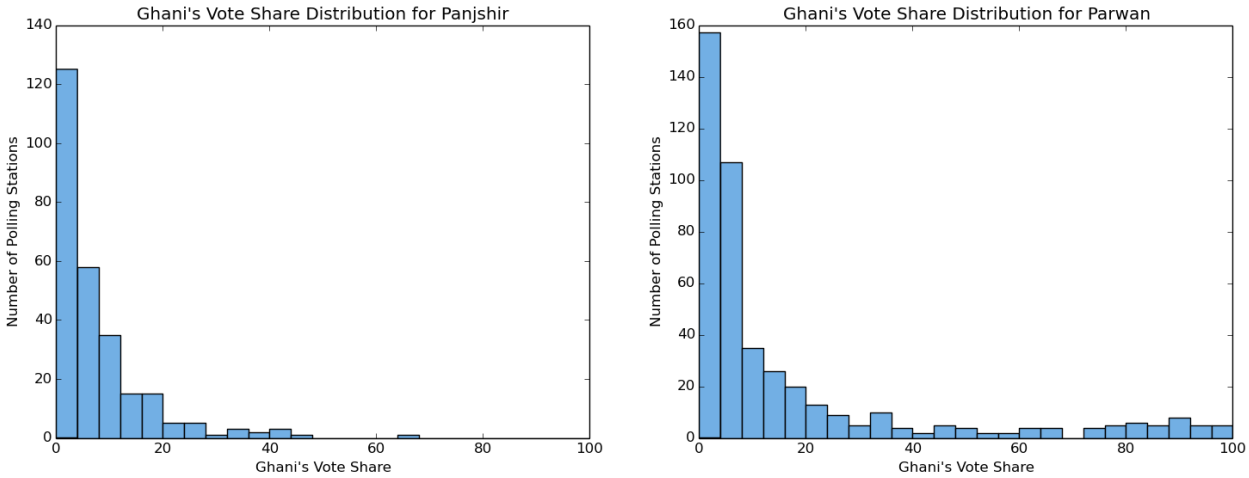
Table 2: *The 20 districts with the highest turnout percentages. Note that all of these districts had higher than or close to 200% turnout, which would mean that their entire CSO-listed population had cast votes in the runoff.*

Clearly, the districts with the highest turnouts were in Paktika, Paktya, Nooristan, Khost, and Kunar. The pro-Ghani province of Paktika is particularly surprising since its districts' turnouts go into the several hundred percent range. While some of these suspiciously high numbers could be explained by pro-Ghani pressures (as discussed in Section 6.1), their magnitudes are so high that we cannot help but conclude ballot stuffing.

6.3 Skewed Vote-Share Distributions

We will conclude our region-specific analysis by looking at the vote-share distributions of suspicious provinces, including some of those in Table 2. To create these plots for a given province, we find the percentage of votes won by Ghani in each polling station there (during the runoff round), and aggregate these vote shares to make a histogram. Note that we focus only on Ghani's vote share in these graphs since Abdullah's vote share distribution is just a mirror image.

Figure 15 shows two examples of vote-share distributions, from the Tajik-dominated provinces of Panjshir and Parwan:



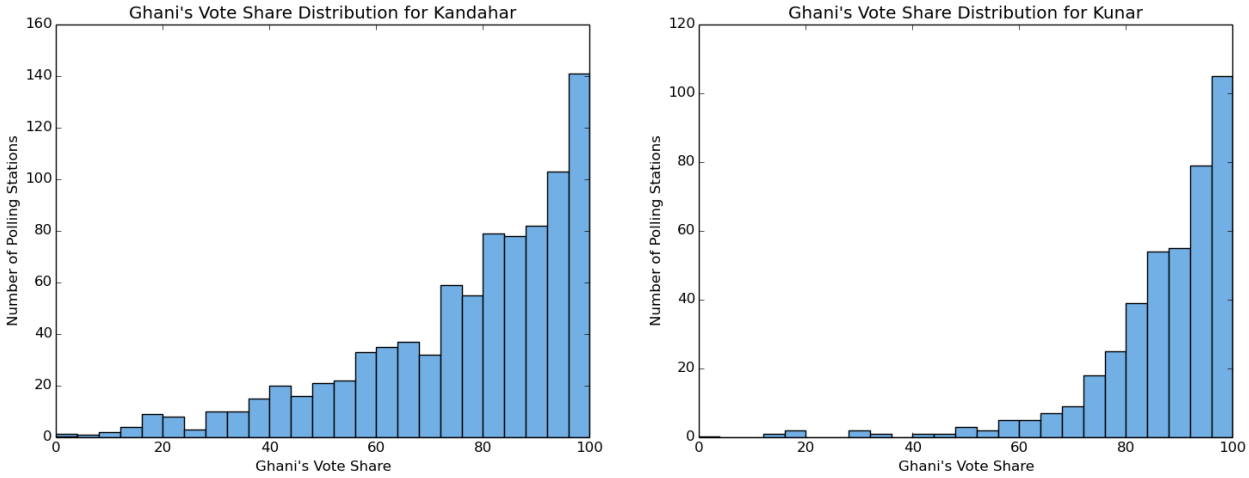
(a) *Ghani's Vote Share Distribution in Panjshir Province* **(b)** *Ghani's Vote Share Distribution in Parwan Province*

Figure 15: *Ghani's vote share in the predominantly Tajik provinces of Panjshir and Parwan (see Figure 1 for the country's ethnic distribution). Given Abdullah's Tajik background and ties to Panjshir, the skews in these figures aren't completely surprising. However, fraud could also be at work.*

It makes sense that these regions would have a vote-share distribution that strongly favors Abdullah over Ghani. After all, as mentioned in Section 2.2.3, Ghani is often associated with the Pashtun ethnicity, while Abdullah – who is half-Pashtun, half-Tajik, and grew up in Panjshir – has a stronger tie with the Tajik and the Panjshiri. Thus, we would expect the Tajik provinces of Panjshir and Parwan to be strongly pro-Abdullah, especially given the importance of ethnicity in the eyes of many Afghan voters [9].

However, we cannot ignore the possibility of fraud in these pro-Abdullah regions, since the skews in Figure 15 are very extreme. Indeed, it is possible that pro-Abdullah election officials were able to stuff ballots and get away with it in these provinces, given how little opposition they would face. Given the high turnout of $\sim 202\%$ in Paryan District in Panjshir (see Table 2), it is possible that some amount of fraud was in fact committed there in favor of Abdullah.

As for Ghani, we would expect the predominantly-Pashtun provinces to vote mostly in favor of him. Based on Figure 16, this does in fact seem to be the case:

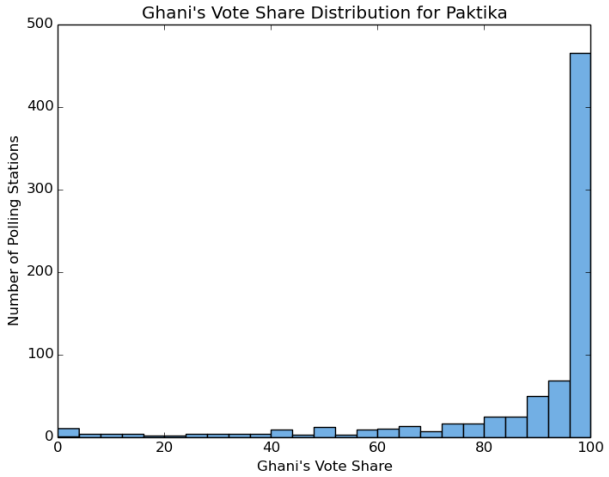


(a) *Ghani's Vote Share Distribution in Kandahar Province* **(b)** *Ghani's Vote Share Distribution in Kunar Province*

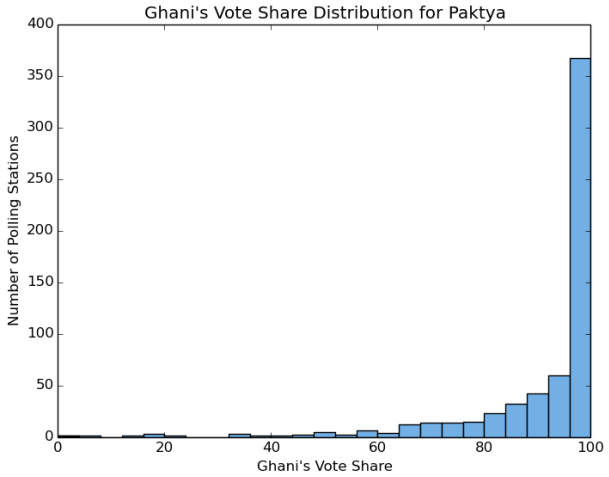
Figure 16: *Ghani's vote share in the predominantly Pashtun provinces of Kandahar and Kunar. Since Ghani is more ethnically Pashtun than Abdullah, it makes sense that he would be favored in these regions.*

One thing worth noting is that – even though Kandahar and Kunar are overall in favor of Ghani – the slightly heterogeneous voter preferences in these regions lead to vote-share distributions that *do not* have an abrupt spike at 100%. Nonetheless, the most common voting share for Ghani in both these provinces is 100%, which is a bit suspicious. Kunar is particularly suspicious since we know (from Table 2) that the district of Asadabad there had a turnout of $\sim 234\%$.

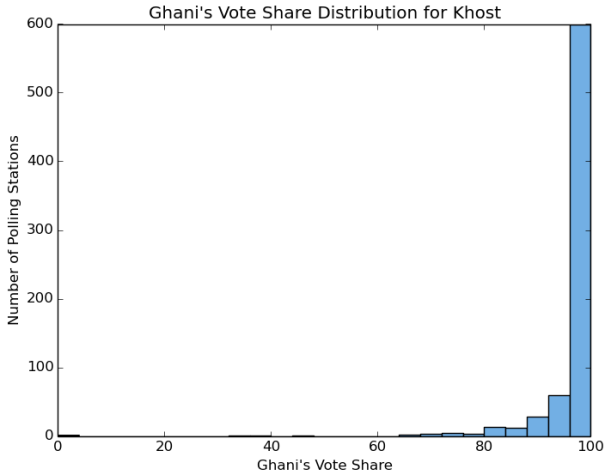
While polling stations in Kandahar and Kunar exhibited a wide range of vote shares in favor of Ghani, the provinces of Paktika, Paktya, Khost, and Wardak had much more suspicious vote-share distributions (somewhat akin to the mirror images of Figures 15a and 15b). These can be seen in Figure 17:



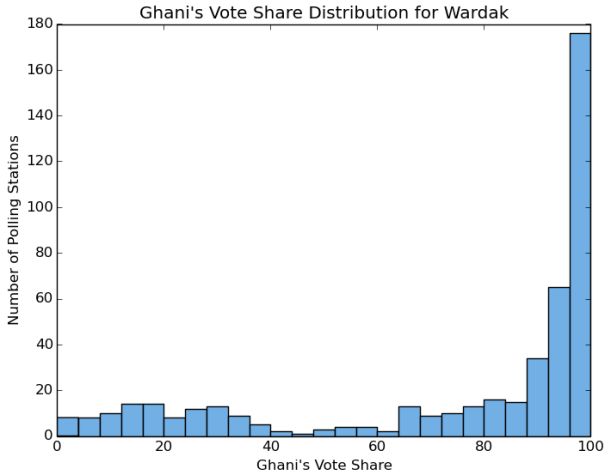
(a) *Ghani's Vote Share Distribution in Paktika Province*



(b) *Ghani's Vote Share Distribution in Paktya Province*



(c) *Ghani's Vote Share Distribution in Khost Province*



(d) *Ghani's Vote Share Distribution in Wardak Province*

Figure 17: *Ghani's extreme vote share in the predominantly Pashtun provinces of Paktika, Paktya, Khost, and Wardak. These histograms are particularly skewed towards 100% vote shares for Ghani, which is highly suspicious – especially given the high turnouts in these provinces.*

Based on these graphs, it appears as if these four provinces have virtually no locations with pro-Abdullah vote shares. While the jump in the count of stations favoring Ghani (at the 100% vote share mark) is dramatic, it could be partly justified by the ethnicity groups in those provinces. Note that 96% of Paktika is Pashtun [27], as is 91% of Paktya [28], 99% of Khost [29], and 70% of Wardak [30]. The strong Pashtun presence in these provinces might have contributed to the high number of polling stations where Ghani won close to 100% of the vote. Also, as discussed earlier, the religious figures in many of these regions preached an obligation to vote for Ghani, and some tribes threatened punishments to those who did otherwise.

However, the possibility of fraud still cannot be ignored, even if we take into account the political impact of Ghani's ethnicity. Many districts in these four provinces produced polling-station-level data that is highly extreme. For instance, as shown in Table 3, Giyan District in Paktika had 34 stations where all 600 ballots (the maximum possible in a polling station) went to Ghani. In fact, Paktika, Paktya, and Wardak all had many such "600 vs. 0" pro-Ghani polling stations:

Province	District	Station Count
Paktika	Giyan	34
Paktika	Urgoon	15
Paktika	Nika	13
Wardak	Chaki Wardak	12
Paktika	Zarghun Shah	11
Paktika	Jaji Aryob	11
Paktika	Sar Hawza	9
Paktya	Laja Ahmad Khail	9
Paktika	Omna	7
Paktya	Janikhel	7

Table 3: *Districts with the highest number of polling stations where all 600 ballots went to Ghani.*

To summarize, the provinces of Paktika, Khost, Paktya, and Wardak exhibit near-exclusive alignment with Ghani. As discussed previously, individuals in these regions were encouraged to vote only for Ghani, and were punished if they did not. Nonetheless, based on the abnormally high turnout rates, lack of observers (in Paktika and Paktya at least, according to Figure 10), and high pro-Ghani vote shares in these four provinces, we have strong evidence that pro-Ghani supporters committed election fraud via ballot stuffing in Paktika, Khost, Paktya, and Wardak.

There is also some evidence for pro-Abdullah strongholds committing fraud. Specifically, the provinces of Parwan and Panjshir are very suspicious, as they both exhibited skewed vote-share distributions and high turnouts – with Panjshir reporting particularly high voting levels (see Table 2). Thus, we have reason to believe that pro-Abdullah supporters also committed large amounts of election fraud via ballot stuffing in these two provinces. However, the scale of Abdullah's vote rigging was nowhere near that of Ghani.

7 Conclusions

We have searched for fraud in the runoff round results of the 2014 Afghan Presidential Election. Below, we reiterate the main results of our analyses in Sections 5 and 6:

1. District-level turnout histograms for both election rounds exhibit clear signs of non-normality, with skewed tails that exceed 100% turnout. In particular, this skew is especially significant in the runoff results. Although poor census data might account for turnout exceeding 100% in some districts, the non-Gaussian behavior of these histograms, coupled with the large number of districts with abnormal turnouts (e.g. 1352% in Wor Mamay, Paktika) suggest ballot stuffing occurred in the runoff election.
2. A province-level winning margin analysis indicates no general (country-wide) correlation between province turnout (minus the average VAP turnout of 50%) and Ghani's winning margin. This suggests that massive, nationwide fraud was not committed by Ghani's side (and *only* Ghani's side) during the runoff election. Nonetheless, we saw signs of localized fraud in the provinces of Khost, Paktika, and Paktya, which had province-level turnouts above 100% as well as very high winning margins for Ghani.
3. Our district-level V/E vs T plots for Abdullah and Ghani do not have particularly strong r^2 values for their linear fits. Nonetheless, they show several districts where $V/E \approx T$ – with many more of these districts voting for Ghani than Abdullah. This suggests that pro-Ghani supporters committed large amounts of ballot-stuffing, while pro-Abdullah voters might have engaged in smaller amounts.
4. Our district-level VS vs T plots for Abdullah and Ghani suggest that vote stealing did not occur in this election. However, these plots *do* show that Ghani was the winner in most "high-turnout" (i.e. $> 100\%$ turnout) districts, which adds further evidence to the ballot-stuffing hypothesis. It's also worth noting that Abdullah did receive large fractions of the vote share in a couple of high-turnout districts – but not as many as Ghani.
5. We visually noted a negative correlation between turnout and observer deployment in the second round, as shown in Figure 10. Furthermore, we found a negative correlation between the relative observer deployment density percent change (RODDPC, measured between the two rounds) and the percent change in turnout between the two rounds. These trends suggest that:

- It was easier to commit fraud in areas with relatively low observer presences.
- The perpetrators of ballot stuffing (in the runoff) may have purposely targeted regions that were relatively less well-observed, compared to the first round.

6. Our first-digit analysis found that Abdullah's district-level first-digit distribution deviated significantly from Benford's Law (in the districts where Abdullah won). However, we cannot really conclude much from this due to aggregation error. It is much more likely that vote falsification occurred at individual polling centers, and we were unable to conduct a Benford's law analysis on polling-station-level data due to the ballot cap of 600 votes per station.
7. Our last-digit analysis plots were not totally uniform for both Abdullah and Ghani (where both plots only looked at the polling stations where the respective candidate won). Yet Abdullah's deviations from the expected uniform distribution were not statistically significant, whereas Ghani's deviations were. Since only the digit zero appeared with a disproportionately high frequency in Ghani's last-digit distribution, we can conclude that pro-Ghani election officials rounded Ghani's vote counts to the nearest tens place, in many of the polling stations that he won.
8. Lastly, based on our region-specific analysis, we have strong evidence that pro-Ghani supporters committed election fraud via ballot stuffing in Paktika, Khost, Paktya, and Wardak. We also have evidence that pro-Abdullah supporters stuffed ballot boxes in Parwan and Panjshir. These conclusions are based on turnouts, vote-share distributions, and (in some cases) low observer deployments.

Based on the above findings, we can conclude that both candidates – Abdullah and Ghani – were likely to have participated in noticeable amounts of regional ballot stuffing, with Ghani committing the vast majority of the stuffing. Moreover, we have support for the hypothesis that election officials falsified vote totals in many of the polling stations where Ghani won.

To detect and/or mitigate future cases of fraud, we suggest several possible corrections. First and foremost, Afghanistan should emphasize updating its census before the upcoming 2019 election. Access to reliable population statistics can be useful in detecting ballot stuffing, and would also be invaluable for day-to-day government affairs. In addition, we propose increasing the number of impartial observers, especially in the aforementioned problematic provinces. This increase in observer deployment must go hand-in-hand with stringent background checks on election officials, since there were many media reports of the IEC engaging in fraud [12]. Lastly, we recommend that polling stations increase the number of available ballots in each province, by either increasing the number of ballots per polling station or increasing the number of polling centers. This correction would address ballot shortage issues that were reported in around half of the provinces [31]. However, the IEC must be careful to closely monitor this increase, since unused ballots in the 2009 election were often used fraudulently and marked exclusively for one candidate [32].

8 Acknowledgments

The authors would like to acknowledge Professor Peter Ordeshook for his insightful discussions and assistance. This paper was completed as part of PS/SS 139, a class on comparative politics and electoral fraud at the California Institute of Technology.

9 Analysis Code and Data

In the spirit of reproducibility, we have included a link to our project's Github repository [33]. This repository not only includes the raw and clean data that we used, but also contains the code we used for generating figures. All programming was done in `python` and `R`.

References

[1] Ritscher, Adam. "A Brief History of Afghanistan." Web. 18 Mar. 2015. URL.

[2] "Mapping the Global Muslim Population." *Pew Research Center*. Pew Research Center, 7 Oct. 2009. Web. 18 Mar. 2015. URL.

[3] "Afghanistan Profile - Timeline." *BBC News*. BBC News, 13 Mar. 2015. Web. 18 Mar. 2015. URL.

[4] "A Historical Timeline of Afghanistan." *PBS Newshour*. PBS, 4 May 2011. Web. 18 Mar. 2015. URL.

[5] "The Constitution of the Islamic Republic of Afghanistan." *The Embassy of Afghanistan*. Web. 18 Mar. 2015. URL.

[6] "Afghanistan Inaugurates New President, CEO in Power-Sharing Deal." *DW*. DW. 29 Sep. 2015. Web. 18 Mar. 2015. URL.

[7] Parvaz, Dorothy. "Afghans Vote in Presidential Poll." *Al Jazeera*. Al Jazeera. 5 Apr. 2014. Web. 18 Mar. 2015. URL.

[8] "Afghanistan Ethnic Groups." *Afghanistan Election Data*. National Democratic Institute. Web. 20 Mar. 2015. URL.

[9] Ibrahimi, Pakteen and Kate Clark. "Elections 2014 (32): A Second Round Surge in Turnout in Loya Paktia?" *Afghanistan Analysts Network*. Afghanistan Analysts Network, 20 June 2014. Web. 23 Mar. 2015. URL.

[10] Graham-Harrison, Emma. "Afghanistan Election Guide: Everything You Need to Know." *The Guardian*. The Guardian, 3 Feb. 2014. Web. 10 Feb. 2015. URL.

[11] Khan, Tahir. "Afghan Elections: Abdullah-Ghani Run-off May 28." *Tribune*. The Express Tribune, 26 Apr. 2014. Web. 20 Mar. 2015. URL.

[12] Gall, Carlotta "In Afghan Election, Signs of Systemic Fraud Cast Doubt on Many Votes." *New York Times*. New York Times, 23 Aug. 2014. Web. 20 Mar. 2015. URL.

[13] Ruttig, Thomas. "Elections (31): Afghanistan's Confusing Election Maths." *Afghanistan Analysts Network*. Afghanistan Analysts Network, 19 June 2014. Web. 23 Mar. 2015. URL.

[14] "Voter Turnout Data for Afghanistan." *International Institute for Democracy and Electoral Assistance*. International IDEA, 25 Aug. 2014. Web. 23 Mar. 2015. URL.

[15] Goldstein, Joseph. "E.U. Confirms Wide Fraud in Afghan Presidential Runoff." *The New York Times*. The New York Times, 16 Dec. 2014. Web. 10 Feb. 2015. URL.

[16] Bijlert, Martine V. "What to Watch in the Elections (2): Polling Station Lists." *Afghanistan Analysts Network*. Afghanistan Analysts Network, 1 Apr. 2014. Web. 25 Mar. 2015. URL.

[17] "2014 Female Only Polling Stations – Breakdown by District." *Afghanistan Open Data Project*. Development Seed. Web. 25 Mar. 2015. URL.

[18] "Afghanistan 2014 Election – About the Data." *Afghanistan Election Data*. National Democratic Institute. Web. 25 Mar. 2015. URL.

[19] Rasmussen, Sune E. "Taliban Return to Afghan Town that Rose Up and Drove Out its Leaders." *The Guardian*. Guardian News and Media, 27 Oct. 2014. Web. 25 Mar. 2015. URL.

[20] "Population." *Central Statistics Organization*. e-Government of Afghanistan. Web. 20 Feb. 2015. URL.

[21] Klimek, Peter et al. "Statistical Detection of Systematic Election Irregularities." *Proceedings of the National Academy of Sciences of the United States of America* 109.41 (2012): 16469-16473. PMC, 24 Sep. 2012. Web. 23 Mar. 2015. doi:10.1073/pnas.1210722109.

[22] Ruttig, Thomas. "Elections 2014 (52): The Not Yet Officially Announced Results – Electoral Maths with Unknowns." *Afghanistan Analysts Network*. Afghanistan Analysts Network, 28 Sep. 2014. Web. 23 Mar. 2015. URL.

[23] "Afghanistan Election Data 2014 – Second Round Election." *Afghanistan Election Data*. National Democratic Institute. Web. 24 Mar. 2015. URL.

[24] Deckert, Joseph, Mikhail Myagkov and Peter C. Ordeshook. "Benford's Law and the Detection of Election Fraud." *Political Analysis* (2011) 19 (3): 245-268. Political Analysis, 2011. Web. 25 Mar. 2015. doi:10.1093/pan/mpr014.

[25] Bezhan, Frud. "Afghan Presidential Transition in Crisis Amid Election Fraud Allegations." *Radio Free Europe*. Radio Liberty, 19 Jun. 2014. Web. 25 Mar. 2015. URL.

[26] Ruttig, Thomas. "AAN Election Blog No. 24: Stuffing and Counting in Paktia." *Afghanistan Analysts Network*. Afghanistan Analysts Network, 24 Aug. 2009. Web. 23 Mar. 2015. URL.

[27] "Paktika Provincial Profile." *Ministry of Rural Rehabilitation and Development*. e-Government of Afghanistan, 2010. Web. 26 Mar. 2015. URL (Internet Archive Link).

[28] "Paktia Executive Summary." *Program for Culture and Conflict Studies*. Naval Postgraduate School, 2010. Web. 26 Mar. 2015. URL.

[29] "Khost Executive Summary." *Program for Culture and Conflict Studies*. Naval Postgraduate School, 2010. Web. 26 Mar. 2015. URL.

[30] "Maydan Wardak Province." *Understanding War*. Institute for the Study of War, 2009. Web. 26 Mar. 2015. URL.

[31] Bezhan, Frud. "Afghan Election: Ballot Shortages Show Lessons Not Learned." *Radio Free Europe*. Radio Liberty, 14 Jun. 2014. Web. 26 Mar. 2015. URL.

[32] Barakzai, Zekria. "An Anti-Fraud Strategy for Afghanistan's 2014 Elections." *United States Institute of Peace*. United States Institute of Peace, 21 May 2013. Web. 26 Mar. 2015. URL.

[33] Bhasin, Laksh and Sharon Yang. "Analysis of 2014 Afghan Presidential Election" *Github Repository*. Github, 28 Mar. 2015. Web. 28 Mar. 2015. URL.