One Person, One Vote: Estimating the Prevalence of Double Voting in U.S. Presidential Elections*

Sharad Goel
Stanford University

Marc Meredith
University of Pennsylvania

Michael Morse
Harvard University and Yale Law School

David Rothschild
Microsoft Research

Houshmand Shirani-Mehr
Stanford University

October 24, 2017

Abstract

There are about three million cases in a national voter file in which 2012 vote records share a common first name, last name, and date of birth. We develop a probabilistic birthdate model to estimate how many of these cases represent the same person voting twice. If registration records are never erroneously marked as being used to vote, we estimate about 0.02% of the votes cast in 2012 were double votes. An audit of poll books, however, suggests that such measurement error could explain many of these apparent double votes. Using data returned to Iowa by the Interstate Crosscheck Program, we quantify the tradeoff between voter accessibility and electoral integrity when purging a likely duplicate registration from another state. We find that one of Crosscheck's proposed purging strategies would eliminate about 300 registrations used to cast a seemingly legitimate vote for every one registration used to cast a double vote.

*We thank Target Smart for supplying us with a national voter file. We thank Delton Daigle, Robert Erikson, Daniel Hopkins, and audience members at the Yale Behavioral Sciences Workshop, the 2017 Midwest Political Science Association Conference, and the 2017 Society for Political Methodology Conference for their comments and suggestions.

1 Introduction

At the heart of the current "voting wars" (Hasen, 2012) lie different beliefs about the incidence of voter fraud (Ansolabehere and Persily, 2008; Stewart III et al., 2016). Such fraud could come in many forms, including voter impersonation, non-citizen voting, or double voting, the last of which occurs when an individual casts multiple ballots, each under a different registration record, in the same election. Levitt (2007) and Minnite (2010) conclude that little to no voter fraud has occurred in recent U.S. elections, largely because there have been few successful prosecutions for such fraud. Not everyone, however, accepts these conclusions, because detecting voter fraud may be difficult when it is done well (Ahlquist et al., 2014). Moreover, while the breadth of evidence brought to bear by Levitt and Minnite is impressive, it is dependent on the investigation of federal, state, or local authorities, or at least on allegations made by them (Fund, 2004, p. 7). Such investigation may be least likely in areas where voter fraud is most likely to be successful.

Because beliefs about the incidence of voter fraud shape the evolving nature of voting rights (Minnite, 2010, p. 129), it is crucial that political science develops new methodologies to quantify and clarify how often it occurs (Alvarez et al., 2009). To this end, we develop a technique to estimate the aggregate amount of double voting, a form of voter fraud that Dick Morris, a prominent conservative political pundit, claimed led to over one million fraudulent votes in the 2012 U.S. presidential election (Morris, 2014). We apply this technique to estimate the number of people voted twice – either by casting two ballots in one state or one ballot in two states – in 2012 using the information contained in a national voter file.

In order to vote in the U.S, a person first must register by providing their name, date of birth, and other personal characteristics. This information, when compiled, is referred to as a voter file, although it would be more accurately described as a voter registration file. For ease, we refer to a voter registration that was marked as having been used to cast a ballot as a vote record.[1] While only one vote can be cast per registration, an individual may be

[1] When we mention the personal characteristics of a particular vote record, we are referring to the personal

registered to vote in multiple jurisdictions. This illustrates an important distinction: a voter is a person, while a vote refers to a specific ballot. Voters who cast a ballot using multiple registrations are double voters.

Identifying double voters is challenging because vote records only consistently report first name, last name, and date of birth (DOB). While 3,047,626 pairs of 2012 vote records in a national voter file share these three attributes, some of these parings represent two distinct voters rather than a double voter. Roughly speaking, we estimate the number of double voters by subtracting the number of distinct voters that we expect to share the same first name, last name, and DOB from the number of observed matching pairs. We build upon McDonald and Levitt's (2008) probabilistic birthdate model for our estimation strategy and extend their work in four ways, accounting for non-uniformity in the distribution of birthdays, producing analytic confidence intervals, explicitly accounting for measurement error in vote history, and looking at the entire country instead of a single state.

If we assume vote records are completely accurate, we would estimate that approximately 33,000 people (0.02% of voters) voted twice in the 2012 presidential election. This implies more than 97% of the possible double votes were cast by distinct individuals. But if roughly 1.3% of registrations were erroneously marked as being used to vote, we would expect to estimate there were around 33,000 double voters even if there were no actual double votes cast. To get a rough sense of the rate of this translation error, we audited a set of poll books in Philadelphia to compare the vote records in the poll books to the constructed vote records in the electronic voter file. This audit detected enough measurement error linking the two sources to explain at least some of the excess double votes.

We validate our statistical model using data generated by the Interstate Crosscheck Program, which compared individual-level voter registration records across participating states to aid in the purging of duplicate records and prosecution of double voters. A participating state receives all pairings in which a registrant in their state and a registrant in another

characteristics associated with the registration record that was marked as having been used to cast a ballot.

participating state share the same first name, last name, and DOB. Crosscheck's data are better than publicly available data at resolving whether two registration records likely belong to the same individual because the data contain, when available, the last four digits of each registration's Social Security number (SSN4). Using data provided to Iowa in 2012, we identified 1,483 pairings with complete SSN4 information in which both registration records were used to vote in 2012. In more than 99.5% of these pairings, the flagged registrations had different SSN4s, supporting our intuition that our model estimates an upper bound on the number of double votes cast in 2012.

Our analysis of Crosscheck data makes explicit the large tradeoff in voter accessibility sacrificed in an administrative environment that only focuses on maintaining electoral integrity. Crosscheck recommends that when it identifies a pairing with the same first name, last name, DOB, and SSN4, a jurisdiction drop the registration record with the earlier registration date. We identified more than 2,500 cases in which only the earlier registration record was used to vote in 2012, compared to just 7 instances in which both the earlier and the later registration were used to vote. Thus, such a strategy would eliminate more than 300 registrations used to cast a seemingly legitimate vote for every double vote prevented.

2 The Measurement of Voter Fraud

Voter-driven fraud, or voter fraud for short, includes the casting of multiple ballots (double voting), illegal ballots (e.g., non-citizen voting), or other people's ballots (voter impersonation). One motivation for the recent wave of photo-identification and proof of citizenship requirements is that a substantial portion of the mass public believes that voter-driven fraud has corrupted the integrity of the electoral process. Between 20% and 35% of the public consistently report that such forms of fraud are very common or occur occasionally (Stewart III et al., 2016). The Supreme Court has voiced concern that such perceptions of fraud "drive honest citizens out of the democratic process and breed distrust of our government"

(*Purcell v. Gonzalez*, 127 S.Ct. 5,7 (2006)). This suggests an important scholarly role for the measurement of voter fraud: if there is little voter fraud, it is particularly important for this to be documented and for the public to be made aware.

Two recent attempts to measure the amount of voter fraud dismiss allegations of rampant cheating of any stripe. Levitt (2007) concludes that "voter fraud is extraordinarily rare," while Minnite (2010) charges that "voter fraud is a politically constructed myth." This work largely relies on what Hood and Gillespie (2012) term "secondary data," which includes election fraud allegations, election incident reports, and journalistic accounts. Both Levitt and Minnite highlight the limited number of criminal prosecutions for voter fraud.

While the rich, anecdotal evidence gathered in Levitt (2007) and Minnite (2010) is critical for lawyers making the case that a specific incident is or is not a violation of election law, it has been dismissed by some who have the prior belief that voter fraud is common. Which election incidents are covered in the first place is dependent on the investigation of state or local authorities, or at least on allegations made by them. This makes it difficult to estimate rates of fraud in the general population based on secondary data, because, as Christensen and Schultz (2013) point out, perhaps only the worst attempts of election fraud leave behind such evidence.

Working under the general umbrella of election forensics, a number of quantitative methods have been developed that overcome some of the limitations of secondary data analysis. Many of these methods examine aggregated data to detect anomalous election outcomes (see, e.g., Cox and Kousser, 1981; Wand et al., 2001). For example, Mebane (2009) tests whether certain numbers show up disproportionately in the second-digit of aggregated election returns, which could happen if vote counters are not fairly tallying ballots. He finds little evidence of such fraud when he applies his test to Florida, Ohio, and nationwide (though see Deckert et al. (2011) for a critique of this Benford-like method and Mebane (2011) for a response). In a similar vein, Cantú and Saiegh (2011) and Montgomery et al. (2015) use machine learning techniques to study improprieties in international elections.

Other methods focus on detecting voter-driven fraud. Ahlquist et al. (2014) use a survey-list experiment to increase the willingness of people reporting that they engaged in voter impersonation and continue to find little evidence of such fraud. Both Hood and Gillespie (2012) and Christensen and Schultz (2013) examine statewide voter rolls and find little evidence of votes being cast using the registration of a deceased individual or statistically anomalous vote histories, respectively.

Little existing election forensics work examines the issue of double voting, despite it being one of the most commonly asserted forms of voter fraud and a factor in the purging of voter rolls (Levitt, 2007). During his tenure as Republican National Committee Chairman, Reince Priebus called double-voters one of the Democrats "core constituencies" (Mayer, 2012). Donald Trump expressed concern during the 2016 presidential campaign that "we may have people vote 10 times" (Weigel, 2016). Dick Morris has gone into more detail about suspected double-voting: "You're talking about probably over a million people that voted twice in [the 2012 presidential election]," he claimed, citing a statement by the Interstate Crosscheck Program as "the first concrete evidence we've ever had of massive voter fraud" (Morris, 2014).

The limited identifiers we observe in voter files make it challenging to identify whether the same person voted in multiple states for the same election. Unlike many other countries, the United States doesn't have national voter registration numbers. While Social Security numbers are sometimes collected and could be used to identify individuals who are registered in multiple jurisdictions, they typically are not shared for privacy reasons. What we can observe are vote records with the same first name, last name, and date of birth (DOB) in multiple states. Although two randomly selected vote records are extremely unlikely to share a common first name, last name, and DOB, a sizable number of matches will occur once we aggregate over the quadrillions of pairs of vote records in the population.[2] While auxiliary

[2]Some states report age or year of birth instead of DOB, which substantially increases the chance of two records sharing common identifiers.

information could be collected to determine whether any given pair represents a case of the same person voting twice or two distinct voters with the same first name, last name, and DOB, doing this at any scale would quickly become time-prohibitive. Instead, we develop a statistical technique to estimate the number of these cases that represent the same person voting twice.

McDonald and Levitt's (2008) study of double voting within New Jersey in 2004 is the best work to date on the topic. McDonald and Levitt identify 884 paris of vote records that share the same first and last name and DOB. They simulate the number of vote records that would be expected to share these observables by drawing the year of birth for a vote record at random from the empirical age distribution of voters and assuming that birthdays within years follow a uniform distribution. Using this method, McDonald and Levitt conclude we should expect 487 cases in which two vote records within New Jersey share the same first name, last name, and DOB, with a 95% confidence interval of plus or minus 102 pairings.

In the sections that follow, we build upon work by McDonald and Levitt (2008) and Yancey (2010) to estimate the number of people who cast two ballots – either in the same state or in two different states – in the 2012 election. Roughly, we estimate the number of double votes by subtracting the number of vote records that we expect to share a common first name, last name, and DOB due to chance from the number we observe in our data. Our model takes a voter's first and last name and year of birth as given and treats their day of birth as a random variable. In addition to expanding the scope of analysis using a national voter file, we deal with two statistical challenges that McDonald and Levitt identify in their approach. First, our model accounts for both name and day of birth periodicity. Second, we also allow for the possibility that some registration records are incorrectly marked as being used to cast a ballot. Failure to account for either issue may cause researchers to overstate the amount of double voting.

Recent work by Enamorado et al. (2017) takes a different approach than the one exemplified by McDonald and Levitt (2008), instead using a Bayesian mixture-model to prob-

abilistically match registrants across voter files. Their approach is part of a broader set of record-linkage models that generate posterior probability estimates that record i from dataset A and record j from dataset B are associated with the same person, even when the number of overlapping variables between the two datasets are inconsistent or some of the variables are measured with error (see also Steorts et al. (2016) and Sadinle (2017)). Additional information can be included when setting the model's priors to better identify matches. For example, Enamorado et al. incorporate information about inter-state mobility between state A and state B into their model when calculating the probability that record i from state A and record j from state B belong to the same person. Thus, these models are well suited for identifying the likelihood that a specific registration in state i and a specific registration in state j belong to the same person.

Our approach, in contrast, is tailored to estimating the aggregate *number* of matches in the population at-large, especially when limited information is observed for each record. Whereas most existing record-linkage models consider only the overall match quality of two fields, we consider the actual values in those fields; as a result, our method can easily account for the varying popularity of names and non-uniform birthdate patterns. Relatedly, most matching approaches evaluate the match quality of two records in a given field independent of the information contained in other fields (though see Sadinle (2017)). In contrast, our models account for interactions that we describe in the next section between someone's first name and date of birth that affect the likelihood that two people who share these characteristics are, in fact, the same person. Such flexibility, in theory, can be incorporated into existing record-linkage methods, but typically comes with significant computational overhead or loss of theoretical guarantees (Enamorado et al., 2017). Our estimation approach, by comparision, scales to hundreds of millions of records while maintaining attractive theoretical properties. While one of the advantages of probabilistic matching models is that data recording errors (e.g., misspelled names) are handled automatically, we address this issue by pre-processing the data to identify and correct such errors before applying our estimation technique.

3 Data

3.1 National Voter File

To estimate the number of people who voted twice in the 2012 election, we use Target Smart's national voter file, which lists the first name, last name, DOB,[3] and turnout history associated with each voter registration. The 124,942,823 vote records in these data provide a nearly comprehensive list of 2012 general election participation.[4]

One of the pitfalls of exact matching is that measurement errors in registration records may influence our estimated rate of double voting. Such bias could point in either direction. An error could eliminate the distinguishing feature between two actually unique vote records, creating the appearance of a double vote, or introduce such a distinguishing feature, masking what would otherwise have been detected as a double vote. As mentioned in Section 2, we attempt to correct both of these issues by pre-processing the data.

Section A.6 in the Appendix highlights a number of data quality issues. Across all years, we found an improbable 14% of 2012 vote records were associated with a first-of-the-month birthday. McDonald (2007) notes that first-of-the-month birthdays are typically used by election officials to identify missing information and drops such "placeholder" birthdates. We follow the same strategy here and remove these records from our baseline analysis that might otherwise cause us to overestimate the number of double votes.

We similarly are concerned that some states have generally poor record-keeping practices, which might introduce an unknown bias into our estimate. Ansolabehere and Hersh (2010) conclude that voter registration data from some states, most notably Mississippi, perform consistently worse than others across a range of data validation exercises. We take advantage

[3]Target Smart sometimes supplements its data with commercial birthdates. We include these observations in our baseline analysis, although we also report results when such observations are dropped.

[4]The FEC reported 129,085,410 presidential votes in 2012. While the vote records most likely to be removed, such as those with a known change of address (McDonald, 2007), also may be used disproportionately to cast double votes, our almost total coverage makes underestimating fraud less of a concern.

of the information contained in both residential addresses and generational suffixes (e.g. "Jr.", "Sr.") to generate two related measures of the accuracy of a state's voter records. We suspect a state may have an in-state duplicate when two records share not only the same first name, last name, and date of birth, but are also registered to vote at the same address. However, differences in the generational suffix suggest that some of these records that look like duplicates may in fact be attributed to measurement error in birthdates. For example, a father and son may share the same first and last name and live in the same household but they should not have the same birthdate if they have different generational suffixes. One should be older. While the vast majority of states have almost no cases of either these related errors, seven states, including Mississippi, have a significant number of them. Because this suggests that there is substantial measurement error in voter records in these states, we drop these states from our baseline analysis.

Finally, we address measurement error in names. Two vote records which should be associated with the same person might not be if each has similar, but not identical, first names. To address this possible underestimation of double votes, the Appendix details how we use commercial software to resolve each first name to its canonical name.

Ultimately, though, this pre-processing approach cannot address all problems of exact matching vote records. For example, while we correct transcription errors in first names, we cannot address the case of outright voter evasion, in which registration records are purposely misleading. That remains a weakness of our inferential approach, though the problem is likely mitigated by established practices of checking registration information against other state databases. To better understand the consequences of measurement error, we present sensitivity analysis in the Appendix that shows how our estimates of the number of double votes are affected by measurement error in birthdates.

Our preferred dataset includes 104,206,280 of the 124,942,823 vote records contained in the full dataset. We explain in the Results section how we adjust our final estimate to account for the dropped records. In doing so, we make an additional assumption that registration

Figure 1: Illustration of a Philadelphia poll book.

records with a first-of-the-month birthday and from the seven dropped states are used to cast double votes at the same rate as all other registration records.

3.2 Philadelphia Voter File Audit

To get a rough sense of the rate at which registrations not used to vote nonetheless have an electronic record of having voted, we compared data on who voted in the 2010 midterm election in Philadelphia, PA according to the poll books to an electronic voter file produced on December 8, 2010. Figure 1 illustrates what a Philadelphia poll book looks like. When people show up to vote in Philadelphia, they sign the poll book next to their name. If someone voted in absentee, this is noted in the signature field. After the election, the voter registration office scans the bar code next to each registration that was used to cast a ballot to generate an electronic record that a vote was cast using that particular registration.

Auditors validated 11,663 electronic registration records with a record of voting and 17,586 electronic registration records without a record of voting in 47 randomly-selected precincts. Our principle interest is in identifying false positives: registrations that had an electronic record of voting, but were not listed as having voted in a poll book.[5]

[5]This might be because the registration record has no corresponding signature in the poll book or because there is no corresponding registration record in that precinct's poll book.

3.3 Iowa Voter File Matched by Crosscheck

The Interstate Crosscheck Program compares individual-level voter registration data across participating states to aid in the purging of double registrations and prosecution of duplicate voters. According to Crosscheck's Participation Guide (see section A.11 in the appendix), participating states upload their voter registration data each January. Kansas administrators return to each participating state a list of registrations in that state that share the same first name, last name, and DOB as a registration in another participating state, which they refer to as "potential double voters." Most of our analysis focuses on 2012, in which Crosscheck handled more than 45 million voter registration records and flagged more than a million potential double voters.

We obtained the list of 100,140 and 139,333 pairings that Crosscheck provided to the Iowa Secretary of State's Office prior to the 2012 and 2014 elections, respectively. In addition to first name, last name and DOB, these data include the middle name, suffix, registration address, county of registration, date of voter registration, voter registration status (i.e., active or inactive), and the last four digits of a registrant's Social Security number (SSN4) in both the Iowa voter file and the voter file of the state of the matched registration. For the Iowa registration, it also includes the voter registration number. For privacy reasons, Iowa removed SSN4 before providing us with these data, instead including an indicator for whether the SSN4 was missing for the Iowa registration, an indicator for whether the SSN4 was missing for the other state's registration, and an indicator for whether the SSN4 was the same in Iowa and the other state.

Knowledge of SSN4 match allows us to better assess whether a specific pairing reported by Crosscheck represents the same individual registered twice or two distinct individuals, each registered a single time. Only 1 in 10,000 distinct people with the same first name, same last name, and same DOB would also share the same SSN4 by chance. So pairings that share all four attributes in common are likely the same person registered twice. And absent transcription error, registrations with different SSN4s are for two distinct people.

To assess the frequency with which votes are cast using the registration records flagged by Crosscheck, we merged the Crosscheck data to the Target Smart national voter file. There are no common identifiers between the two datasets for all participating states, so we exactly matched records on first name, middle name, last name, date of birth, and state.[6] Because our Target Smart data were generated after our Crosscheck data, a registrant's information may have changed between when Crosscheck identified its pairings and when the Target Smart data were compiled. In addition, some of the information reported to Crosscheck may not have been reported to Target Smart, particularly if such information is privileged or confidential. Because we are concerned that some registrants in Crosscheck will fail to match to their own vote record in Target Smart,[7] we also merged the Iowa-specific registration records flagged by Crosscheck with a contemporaneous Iowa voter file using the voter registration number that is contained in both sources.

4 Analysis

4.1 Model

To derive a statistical estimator for the number of people k who voted twice in a given election, we first decompose k into the sum $k = \sum_f \sum_l \sum_y k_{f,l,y}$, where $k_{f,l,y}$ is the number of double voters with first name f and last name l who were born in year y. While we cannot observe $k_{f,l,y}$, we can estimate it by combining three quantities. The first is $N_{f,l,y}$: the number of vote records in a given election with first name f, last name l, and birth year y. The second is $M_{f,l,y}$: among the $N_{f,l,y}$ vote records described above, $M_{f,l,y}$ is the number of pairs of records having the same birthday b. Finally, we consider $p_{b|f,l,y}$: the probability of

[6] It is appropriate to use middle name in this match because we assume that we are generally matching to the exact registration record identified by Crosscheck.

[7] A hand-coding of Crosscheck records that failed to match to the Target Smart voter file suggests that there are a few people in the Crosscheck data who voted, but failed to match to the Target Smart voter file, because of discrepancies in the date of birth between the two sources.

having a birthday b conditional on having first name f, last name l, and being born in year y.[8] Whereas we can directly observe the first two quantities from the voter file, we must estimate $p_{b|f,l,y}$, as described below. The intuition behind our approach to estimating double voters is that we subtract from $M_{f,l,y}$ the number of vote records that we would expect to share the same birthday given $N_{f,l,y}$ and $p_{b|f,l,y}$.

We make three key assumptions in order to estimate the number of double votes cast. First, we assume that the voter file is a completely accurate reporting of whether a registration was used to vote in a given election. When this assumption holds, double voting is the only explanation for why the same individual would be recorded as having voted in two states. We revisit this assumption in the next section, when we investigate the effect of recording errors on our estimate.

Second, we assume that an individual votes at most twice. Pew (2012) finds that very few people are registered to vote in more than two states. Moreover, among all cases in our data of vote records matching on name and date of birth, only 6% involve matches of three or more records.

Finally, we assume that the birthday distribution $p_{b|f,l,y}$ can be well approximated as follows. Define $d_{b,y}$ as the day of the week on which birthday b occurred in year y. For instance, $d_{\text{September 25, 1970}} = \text{Friday}$. Next, define $C_{f,y,b}$ as the number of voters with first name f, birthday b, and birth year y. Let B, F, and D be random variables that specify the birthday, first name, and birth day-of-week of a random voter. Then we estimate $p_{b|f,l,y}$ by

$$\hat{p}_{b|f,l,y} = \frac{\hat{\Pr}(B = b \mid F = f)\, \hat{\Pr}(D = d_{b,y})}{\sum_{b'} \hat{\Pr}(B = b' \mid F = f)\, \hat{\Pr}(D = d_{b',y})}. \tag{1}$$

The factors in Eq. (1) are given by

$$\hat{\Pr}(D = d) = \frac{\sum_{f'} \sum_{y'} \sum_{b'} C_{f',y',b'} \mathbb{1}(d_{b',y'} = d)}{\sum_{f'} \sum_{y'} \sum_{b'} C_{f',y',b'}} \tag{2}$$

[8] $p_{b|f,l,y}$ is shorthand for $\Pr(B = b \mid F = f, L = l, Y = y)$

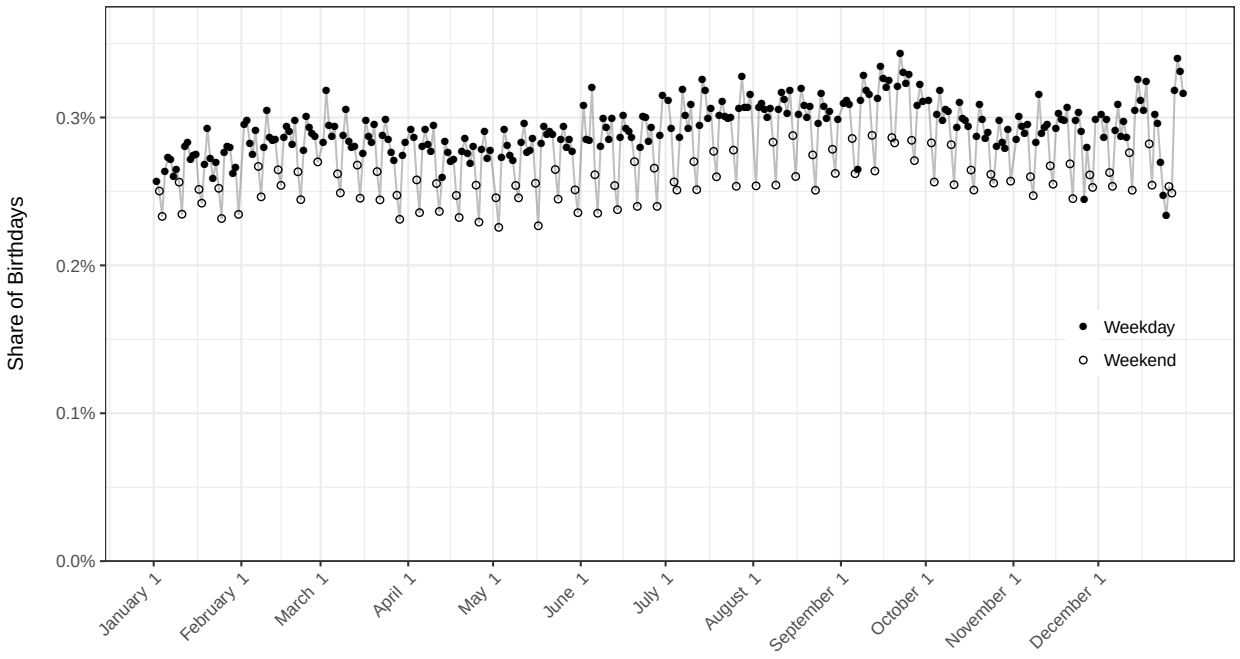


Figure 2: Distribution of (cleaned) birthdays in 1970 in the voter file.

and, for a smoothing parameter $\theta = 10,200$ that maximizes model fit,[9] we set

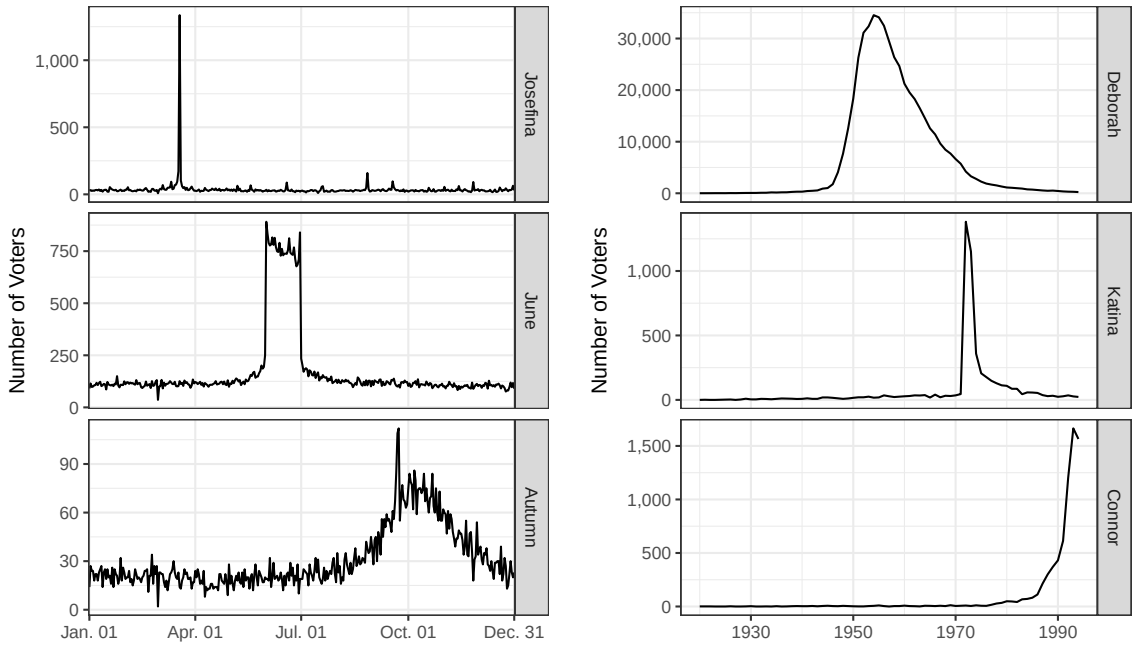
$$\hat{\Pr}(B = b \mid F = f) = \frac{\theta \hat{\Pr}(B = b) + \sum_{y' \neq y} C_{f,y',b}}{\sum_{b'} (\theta \hat{\Pr}(B = b') + \sum_{y' \neq y} C_{f,y',b'})} \tag{3}$$

where

$$\hat{\Pr}(B = b) = \frac{\sum_{f'} \sum_{y'} C_{f',y',b}}{\sum_{b'} \sum_{f'} \sum_{y'} C_{f',y',b'}}. \tag{4}$$

Our estimates of $\Pr(D = d)$ and $\Pr(B = b)$ in Eqs. (2) and (4) aggregate over all voters to generate the empirical distributions. Our estimate of $\Pr(B = b \mid F = f)$ in Eq. (3) averages the birthday distribution specific to each first name f with the overall distribution aggregated over all first names in every year, excluding observations from year y to remove the effect of a specific registrant's own data when estimating the probability that he or she was born on a given day.

[9]This θ maximizes the likelihood of observing the data under the model, as estimated on a random 1% sample of vote records held out when constructing $\hat{p}_{b|f,l,y}$.



(a) (b)

Figure 3: Examples of names among 2012 voters with a non-uniform date of birth distribution, by day (a) or year (b) of birth.

Section A.1 presents the assumptions that justify this modeling strategy. At a high-level, the estimator $\hat{p}$ couples periodicities in birth day-of-week with seasonal correlation between first names and birthdays. Using the data on 2012 voters born in 1970, Figure 2 illustrates that the same number of people are not born on all days. For example, people are more likely to be born during the summer than during other parts of the year and on weekdays than on weekends. Figure 3 also shows that certain first names are more frequently observed among people born in certain points of the year and in certain years. A disproportionate number of voters named June were, unsurprisingly, born in June, while voters named Katrina were more likely to be born in the 1970s than either the 1960s or 1980s.

Figure 4 shows the modeled distribution of birthdays of voters born in 1970 for five different first names, and how they compare to the empirical distribution of birthdays. The names in the plot are ordered from top to bottom based on their popularity among voters. For names like Michael, which have a mostly uniform birthday distribution in a year, our model

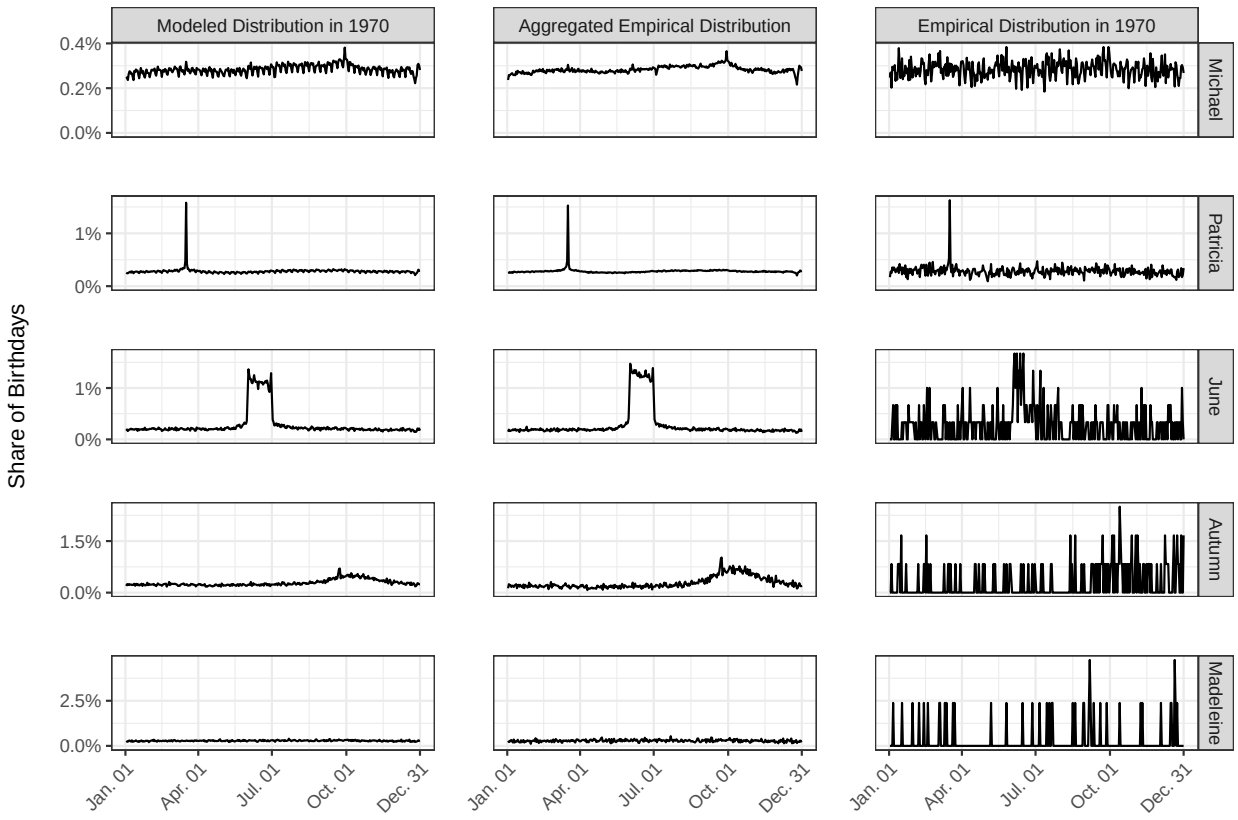


Figure 4: Modeled distribution of birthdays for voters born in 1970 for five different first names vs. the empirical distribution of birthdays for votes with those first names (aggregated across all years) and the empirical distribution of birthdays for voters with those first names born in 1970. Across all years (in 1970), we observe 1,669,641 (39,583) voters named Michael, 894,836 (7,621) voters named Patricia, 60,464 (299) voters named June, 10,956 (120) voters named Autumn, and 7,084 (42) voters named Madeleine.

captures day-of-week and seasonal effects well. Additionally, for names with non-uniform birthday distributions and different levels of popularity, like Patricia, June, or Autumn, our method is able to capture the cyclic popularity of the first names. Finally, for highly infrequent names, like Madeleine, are model captures only aggregate, non-name-specific day-of-week and seasonality trends.

We combine these estimates of $p_{b|f,l,y}$ with knowledge of $M_{f,l,y}$ and $N_{f,l,y}$ to estimate $k_{f,l,y}$ using the following theorem.

Theorem 4.1 *Suppose $D_{f,l,y}$ is a discrete probability distribution of birthdays $b_1, \ldots, b_n$ with*

$\Pr_{D_{f,l,y}}(b_i) = p_{b_i|f,l,y}$. *Further assume there are* $m \geq 1$ *independent observations from* $D_{f,l,y}$, $B_1, \ldots, B_m$, *and* $k_{f,l,y} \leq m$ *copies* $B_{m+1}, \ldots, B_{m+k_{f,l,y}}$ *such that* $B_{m+i} = B_i$. *Let* $M_{f,l,y}$ *be the number of pairwise matches among the* $N_{f,l,y} = m + k_{f,l,y}$ *observations, and define the estimator*

$$\hat{k}_{f,l,y} = \left(M_{f,l,y} - \binom{N_{f,l,y}}{2} \sum_i p_{b_i|f,l,y}^2 \right) \Bigg/ \left(1 - \sum_i p_{b_i|f,l,y}^2 \right). \tag{5}$$

Then $\mathbb{E}\hat{k}_{f,l,y} = k_{f,l,y}$ *and*

$$\mathrm{Var}(\hat{k}_{f,l,y}) \;\leq\; 4\binom{N_{f,l,y}}{2} \left[\frac{\sum_i p_{b_i|f,l,y}^2}{1 - \sum_i p_{b_i|f,l,y}^2} \right] + 12\binom{N_{f,l,y}}{3} \left[\frac{\sum_i p_{b_i|f,l,y}^3 - \left(\sum_i p_{b_i|f,l,y}^2 \right)^2}{\left(1 - \sum_i p_{b_i|f,l,y}^2 \right)^2} \right].$$

The proof for Theorem 4.1 is included in the Appendix.

Section A.5 describes a simulation we use to investigate the finite sample properties of our model, taking into account sampling error introduced by estimating the probabilities $p_{b|f,l,y}$ from data. We generated 100 synthetic datasets that maintained a similar correlational structure as the true data between first name, birth year, and birthday conditional on first name and birth year. In each synthetic dataset, we copied a known number of records to simulate the presence of somewhere between 0 and 45,000 double votes. We then use our model to estimate how many records we copied. Figure A.2 demonstrates that our model does a good job of estimating the number of copied observations, and that our analytic standard errors are, if anything, slightly too conservative.

As discussed in the Data section, we believe that the birthdates of some observations are particularly likely to be measured with error. Specifically, we are concerned about observations with a first-of-the-month birthday and observations from states that appear to misreport the birthdates of individuals in multi-generational households. To account for this, we next present a lemma that allows us to estimate the rate of double voting in the population after dropping observations that we suspect are the most likely to be inaccurate. We also build in the possibility of duplicated records, which we use in the next section to

incorporate measurement error in whether a registration record was used to cast a ballot.

To derive this new estimate, we assume voter registrations go through a stochastic update process in which each record is duplicated with probability p_u and dropped with probability p_r. Lemma 4.1 estimates the original number of double voters before the update happened, $k^{\text{orig}}_{f,l,y}$, based on the number of double voters that end up in the updated sample, $K_{f,l,y}$. Though we cannot directly observe $K_{f,l,y}$, we can estimate it using Theorem 4.1.

Lemma 4.1 *Assume a set of $n \geq 1$ objects, out of which $k^{orig}_{f,l,y}$ objects are duplicates, and the rest are unique. Additionally assume that each object has at most one duplicate in the set. Then suppose that each one of these n objects is copied with probability p_u, and dropped from the set with probability p_r. Assume $K_{f,l,y}$ to be the number of unique objects with a copy in the updated set, and $N_{f,l,y}$ to be the size of this set. If we define the estimator $\hat{k}^{orig}_{f,l,y}$ as,*

$$\hat{k}^{orig}_{f,l,y} = \frac{K_{f,l,y}}{(1-p_r)^2 - 2p_u} - \frac{N_{f,l,y}p_u}{(1+p_u-p_r+p_up_r)\left((1-p_r)^2-2p_u\right)} \tag{6}$$

then $\mathbb{E}\hat{k}^{orig}_{f,l,y} = k^{orig}_{f,l,y}$.

The proof for Lemma 4.1 is included in the Appendix. Based on the lemma, if we have $p_r \gg p_u$ then the estimator simplifies to,

$$\hat{k}^{\text{orig}}_{f,l,y} \approx \frac{K_{f,l,y}}{1-2p_r} - \frac{N_{f,l,y}p_u}{(1-p_r)(1-2p_u)}. \tag{7}$$

4.2 Results

We first exclude observations with data quality issues, as described in Section 3.1, to obtain our preferred sample of just over 104 million vote records. Within our preferred sample, there are 761,875 pairs of 2012 vote records that share the same first name, last name, and DOB. Given our assumptions about $p_{b|f,l,y}$, we next estimate that there were 21,441 (s.e. = 1,727) double votes using Theorem 4.1. Using Lemma 4.1, we scale the results of our analysis on our preferred sample to account for the observations we excluded. Given

that the FEC reported that just over 129 million votes were cast in the 2012 presidential election, a vote record has about a 19.3% chance of being dropped from our preferred sample. Lemma 4.1 says we can generalize the rate of double voting in the broader population by multiplying the estimated number of double voters in our preferred sample by 1.53. Thus, we estimate there were 32,890 (s.e. = 2,649) double voters in the population.

Tables A.3 and A.4 demonstrate the sensitivity of our results to a number of the assumptions we make in our analysis. Table A.3 focuses on sample restrictions, and shows that the number of estimated double votes would be somewhat higher if we kept observations from states with high rates of duplicate voting for people living in the same address, unaffected by the presence of observations with a commercially sourced date of birth, and substantially higher if we didn't exclude observations with a first-of-the-month birthday.[10] Using our preferred birthday distribution reduces the estimated number of double votes in our preferred sample by just over 25%. A little under half of this reduction results from accounting for periodicity that affects all first names and a little over half of this reduction results from accounting for name-specific periodicity. Table A.4 shows that our results are not particularly sensitive to the standardization of first name and assumptions about the smoothing parameter θ in our birthday distribution function. We also find that the standard error we estimate from the bootstrapping procedure we describe in Section A.5 in the Appendix is less than our analytic standard error, which makes sense given that our analytic standard error is an upper bound on sampling error.[11]

[10] We specified Equations 1, 2, and 3 knowing that observations with first-of-the-month birthdays would be dropped in our baseline model. While our estimates are sensitive to the inclusion of these observations, we would specify these equations differently to better account for the excess number of people with a first-of-the-month birthday if we were trying to estimate the number of double votes with these observations remaining in the sample.

[11] We also show in this section that a one percent rate of measurement error in birthdates causes us to underestimate the number double votes by 2.2 to 2.5 percent.

4.3 Accounting for Measurement Error in Vote Records

The results presented in the previous subsection assumed that vote records in the voter file are a completely accurate representation of who did and did not vote in a given election. But that assumption may not be realistic given what Minnite (2010, p. 7) describes as a "fragmented, inefficient, inequitable, complicated, and overly complex electoral process run on Election Day essentially by an army of volunteers." Such a complex process creates many points where human error could result in the voter file inaccurately indicating whether a particular registration record was or was not used to vote. This is only compounded by the problem of deadwood registration. The remainder of this subsection extends the model from the previous section to incorporate the possibility that registrations not used to vote nonetheless have a vote erroneously recorded in the voter file. We account only for the measurement error in vote records associated with cross-state duplicate registrations because we believe that it is substantially easier for a state to drop duplicate records within a state than across states.[12] The following subsection then presents evidence from an audit of Philadelphia poll books that is useful for assessing the rate at which these errors may happen.

A thought experiment illustrates how even minor errors in the recording of votes in a voter file could generate a substantial number of cases of illusory double voting. Imagine a world with 140 million registration records, 100 million of which were used to cast a ballot in an election. If a vote record is mistakenly attached to a non-voting registration in 1% of the cases, this would result in one million records, or 2.5% of non-voting registrations, being incorrectly marked as being used to cast a ballot. Some number of these registration records are dormant deadwood registrations of people who moved to, and voted in, a different jurisdiction. Assuming recording errors are assigned randomly, we would generate 2,500

[12]Figure A.3 suggests there are few duplicate records per capita among the states in our baseline analysis. But to the extent that there are such duplicates, our revised estimate will be conservative, since it will not account for this type of measurement error.

illusory double votes for every 100,000 voters that have a deadwood registration.[13]

We can use Lemma 4.1 to account for this potential measurement error in the recording of votes. Let f_p represent the probability of a false positive, such that a registration record that was not used to cast a ballot nonetheless has a vote record associated with it. Similarly, let f_n represent the probability of a false negative, such that a registration record that was used to cast a ballot does not have a vote record associated with it. Additionally, let c be the number of cases where a voter has a duplicate registration record in another state, and N be the total number of votes in the election. In the context of Lemma 4.1, p_u represents the probability of a vote record getting duplicated in the voter file, which corresponds to cases where a deadwood registration for a voter in another state is wrongly recorded as having voted. We can model $p_u = \frac{c(1-f_n)f_p}{N}$.[14] Further, p_r is the probability of a voter record getting dropped, which is the same as the false negative rate, and so $p_r = f_n$. Finally, $K_{f,l,y}$ represents the number of double voters we observe after measurement errors are accounted for, which is estimated from Theorem 4.1.

Evaluating this model requires an estimate of the number of duplicate registrations for voters (c) as well as the probability of observing false positive (f_p) and false negative (f_n) vote records in a voter file. We follow a procedure similar to the one derived in Theorem 4.1 to estimate the number of deadwood registrations for voters. While we cannot observe c directly, we can compute T, the number of observed cases in which two registration records in different states share the same first name, last name, and date of birth, and exactly one of them is recorded as having voted in the given election. As before, the estimator approximately subtracts from T the number of cases we would expect to observe due to chance in which a vote record and a non-voting registration record in different states share

[13] Pew, 2012 reports that about 2.75 million people are registered to vote in multiple states.

[14] Assuming c voters have a duplicate registration record in another state, we can estimate $c(1-f_n)$ of them to have their votes correctly recorded. Out of the duplicate registration records for these $c(1-f_n)$ voters, we expect $c(1-f_n)f_p$ of them to be incorrectly recorded as voted. Therefore, the ratio of voters that are duplicated due to measurement errors is $\frac{c(1-f_n)f_p}{N}$.

the same first name, last name, year of birth, and birthday given our estimates of $p_{b|f,l,y}$.

Our estimate involves four key assumptions. First, as before, we assume that registration records are fully accurate. Second, we assume that each individual is at most registered in two states.Third, we assume that our estimate of the birthday distribution, modeled as before, is accurate. Lastly, we assume individuals are listed in the poll books for a state if they have voted in that state in at least one of two previous elections.[15]

We define $c = \sum_f \sum_l \sum_y c_{f,l,y}$, where $c_{f,l,y}$ is the number of voters with first name f, last name l, and year of birth y who have a duplicate registration record in another state. Denote by $B_1, \ldots, B_m$ the birthdays for unique registration records with first name f, last name l, and birth year y. We assume these observed birthdays are $m \geq 1$ samples from a discrete probability distribution $D_{f,l,y}$ with values $b_1, \ldots, b_n$ and $\Pr_{D_{f,l,y}}(b) = p_{b|f,l,y}$. We further assume each of these registration records corresponds to one of u states we are analyzing named $S_1, \ldots, S_u$. We can enter cross-state duplicate registrations into our framework by assuming $0 \leq k \leq m$ duplicate records with birthdays $B_{m+1}, \ldots, B_{m+k}$ which are generated as $B_{m+i} = B_i$, and are scattered in $S_1, \ldots, S_u$. Finally, indicate whether observation B_i for $1 \leq i \leq m + k$ has been recorded as voted or not by a flag f_i. In terms of this notation, $c_{f,l,y}$ is the number of duplicate pairs $\{(B_i, B_{i+m}) \mid 1 \leq i \leq k\}$ such that exactly one of the elements of the pair has voted, and $T_{f,l,y}$ is the number of pairwise matches among $m + k$ observations such that the two elements of the pair are from different states and exactly one of them has voted. Theorem 4.2 provides an estimator for $c_{f,l,y}$ based on $T_{f,l,y}$.

Theorem 4.2 *Let v_l be the number of observations that voted in state S_l ($v_l = \sum_{B_i \in S_l} f_i$), and $\bar{v}_l$ the number of observations without a vote in that state ($\bar{v}_l = \sum_{X_i \in S_l} (1 - f_i)$). Define the estimator*

$$\hat{c}_{f,l,y} = \left(T_{f,l,y} - \left(\sum_{l=1}^{u} v_l \sum_{l=1}^{u} \bar{v}_l - \sum_{l=1}^{u} v_l \bar{v}_l \right) \sum_i p_{b_i|f,l,y}^2 \right) \Bigg/ \left(1 - \sum_i p_{b_i|f,l,y}^2 \right). \qquad (8)$$

[15]Pew (2012) uses data compiled by the U.S. Election Assistance Commission to report that failure to vote in two consecutive federal elections is the modal reason someone was removed from a state voter file.

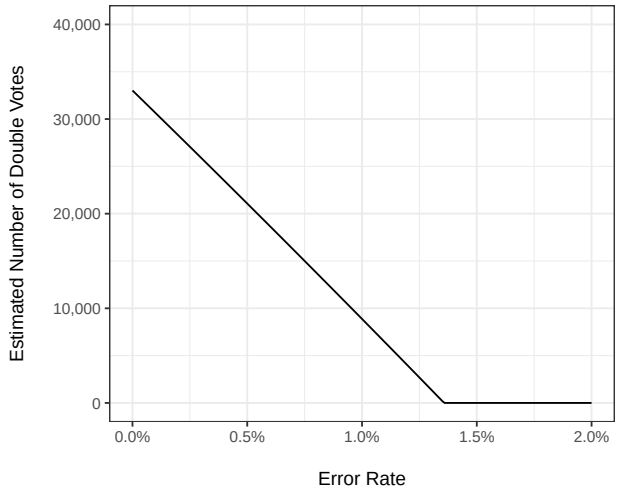


Figure 5: How the estimated number of double votes changes based on the clerical error rate.

Then $\mathbb{E}\hat{c}_{f,l,y} = c_{f,l,y}$ *and*

$$\mathrm{Var}(\hat{c}_{f,l,y}) \leq \left(\sum_{l=1}^{u} v_l \sum_{l=1}^{u} \bar{v}_l - \sum_{l=1}^{u} v_l \bar{v}_l \right)^2 \left[\frac{\sum_i p_{b_i|f,l,y}^2}{1 - \sum_i p_{b_i|f,l,y}^2} \right]. \tag{9}$$

The proof for Theorem 4.2 is included in the Appendix.

We observe $T = 1,818,666$ pairs in the voter file, and the theorem estimates that 1,579,592 (s.e. = 22,186) of these are duplicate registrations. If we assume $f_p = f_n = f$ as the clerical error rate, then we should plug in $p_u = \frac{1,579,592 \times f \times (1-f)}{104,206,280}$ and $p_r = f$ in Lemma 4.1 to correct for measurement error. After scaling the corrected estimate of double votes for records we dropped from our analysis using the steps discussed in the previous subsection, Figure 5 shows how the point estimate of the number of double votes changes with respect to different error rates. It illustrates that a clerical error rate of about 1.3% would be sufficient to explain all of the double votes we estimate.

Our Philadelphia poll book audit data show that 1% of registrations without a vote record in the poll book nonetheless have an electronic vote record. This suggests that, at a minimum, our estimated number of double votes is an over-estimate. If our Philadelphia

audit were representative of the false-positive rate in the population, Figure 5 indicates that our estimate would drop to about 10,000. These audit results, though, are only meant to be illustrative of the false positive rate in the population.

Multiple notes of caution are discussed in more detail in Section A.7 in the Appendix. The false positive rate in Philadelphia may be larger than the rate in the general population, perhaps because Pennsylvania is known to have more voter file discrepancies (Ansolabehere and Hersh, 2010), but it may also be smaller, since the local office has a large, professionalized, and experienced staff. Further, while a small, but growing number of jurisdictions use an electronic poll book to update vote history, Philadelphia's poll-book-and-bar-code approach likely produces fewer errors than a sign-in sheet with no bar codes, which requires manual entry. Also there are forms of false positives that our audit would not detect. For example, Hopkins et al. (2017) report that 105 individuals were forced to file a provisional ballot in a state election because their registration was wrongly marked as having been used to vote earlier in the day.

Because we our knowledge of the false positive rate in the population is so rough, we cannot say anything definitive about how many of the potential double votes can be explained by measurement error. Ultimately, all we can conclude is that measurement error likely explains a sizable portion, and possibly nearly all, of the surplus double votes that we observe in the national voter file.

4.4 Model Validation

We conclude the paper by validating our estimated number of double votes in Target Smart's national voter file using the ratio of likely double votes to potential double votes observed between Iowa and other states participating in the Interstate Crosscheck Program. As discussed earlier, a subset of the data provided by Crosscheck to Iowa prior to the 2012 and 2014 elections has complete SSN4 information for both the Iowa registration record and the registration record in another participating state with the same first name, last name,

Table 1: Vote Records among Registration Pairings Returned to Iowa by Crosscheck

Year of Data	2012		2014	
SSN4 Match	Yes	No	Yes	No
# of Registrations (Reg.)	25,987	8,913	34,189	14,766
Which Reg. Used to Vote:				
Both	7	1,476	9	2,809
One (earlier reg. date)	2,543	1,687	2,019	3,430
One (later or unknown reg. date)	9,429	2,572	8,612	2,697
Neither	14,008	3,178	2,3549	5,830

and date of birth. We match these data to turnout data to identify a set of potential double votes. When these registration records share the same SSN4, a potential double vote is likely an actual double vote, subject to the caveats about measurement error discussed in the previous section. In contrast, when two registration records have different SSN4s, they likely represent records for two distinct individuals who just happen to share the same first name, last name, and date of birth.

Table 1 shows that the incidence of actual double votes is extremely low in the registration pairings identified by Crosscheck based on first name, last name, and DOB. (The table only includes pairings in which SSN4 information is available for both records in the match.) There are seven cases in 2012 in which registration records with the same SSN4 were both used to cast a vote. In contrast, there were 1,476 cases in which registration records with different SSN4s were both used to vote. Thus, the probability of a registration pairing sharing the same last four digits of a social security number conditional on both registrations being used to cast a ballot was about 1 in 200 in 2012. The same quantity in 2014 was about 1 in 300.[16]

In our preferred sample, there were 761,875 pairs of 2012 vote records that share the same

[16]One concern is that measurement error in turnout may obscure some double voting, either because of differences in the measurement of date of birth in the Target Smart and Crosscheck data or because voter records are missing in the Target Smart data. But Section A.8 in the Appendix shows very similar patterns in 2012 if turnout in Iowa is directly measured from the voter file and we restrict the analysis to states in which fewer than 10% of vote records have a birthday on the first of the month.

first name, last name, and date of birth. If only 1 in 200 of these records is an actual double vote, there would only be about 3,800 double votes, as opposed to our estimate of about 21,000. We believe, though, that the actual number of double votes within this population probably falls somewhere in between these two estimates. There are a number of reasons why these estimates might differ. First, the Crosscheck data focus on at inter-state double voting, while our model also capture intra-state double voting. Our model estimate may be too high, for example, if there are additional sources of birthday homophily that our conditional birthday probability model fails to capture. It could also be that more than 1 out of every 200 potential double votes in the population is a likely double vote. There is some measurement error in vote history that may cause some registration records used to cast a ballot to not have a vote record attached to it. Further, the ratio was estimated using potential double votes from a non-random subset of states—those in Crosscheck—and a non-random subset of registrants within those states—those with known SSN4s.

The latter is particularly important to note when considering the policy implications of a Crosscheck-style program. About two-thirds of potential double registrations identified by Crosscheck have at least one unknown SSN4. Thus, more often than not, an election administrator will not have enough information to distinguish good and bad matches. Based on the subset of pairings with SSN4 data, one in four potential double registrations returned by Crosscheck are likely not actual double registrations; if SSN4 information is missing at random, this ratio should also hold in the full set of returned matches.

Election administrators may believe that the aggregate match quality is sufficiently high to justify dropping the registration with the earliest registration date in each matched pair. Table 1 shows that more than 1,000 legitimate votes would be impeded for every double vote prevented by this strategy. We describe such votes as being impeded rather than denied, because some of the affected people would recognize that their registration was at risk of being removed and take the necessary steps to fix the situation before Election Day. But others likely would not.

In 65% of the cases, the dropped registration would have been the product of a bad match, an instance where two people share a first name, last name, and DOB by chance. To be fair, the Crosscheck guide does not recommend purging the earlier registrations without a confirmation of SSN4. However, the guide does recommend purging when a potential double registration does share a common SSN4. But problems remain even when it is known two registration records belong to the same person. While the later registration record is more likely to be used to cast a vote, the registration record with an earlier registration date was used to cast a ballot 2,543 times. Thus, purging the record with the earlier registration date would risk impeding more than 300 votes for every double vote prevented.

5 Discussion

The evidence compiled in this paper suggests that double voting is not carried out in such a systematic way that it presents a threat to the integrity of American elections. In an election in which about 129 million votes were cast, we estimate that at most only 33,000, or 0.02 percent of votes cast, were double votes. This estimate should be taken as an upper bound on the amount of double voting, as our Philadelphia audit suggests many of these apparent double votes could be a result of measurement error in turnout records. Thus, there is almost no chance that double votes could affect the outcome of a national election.

These findings may come as a surprise to a number of Americans who report on surveys that double voting is not rare. Stewart III et al. (2016) find that about 25% of the public believes that voting more than once happens either commonly or occasionally (as opposed to infrequently or never), while another 20% report being unsure how often it happens. Such beliefs are likely driven, at least in part, by media stories that conflate the number of observed cases of voter records sharing the same observable characteristics with double votes. For example, Johnson (2014) used the headline "N.C. State Board Finds More than 35K Incidents of 'Double Voting' in 2012" to describe Crosscheck's finding that the 35,750

North Carolina vote records from 2012 shared the same first name, last name, and date of birth with registrations used to vote in other states.

Such media stories demonstrate the importance of clearly differentiating between (1) registration records that share common observable characteristics, (2) duplicate registrations, and (3) double votes. In 2013, Crosscheck circulated Figure A.6 which reported that it had identified 1,395,074 "potential duplicate voters"—registration records with a common first and last name and date of birth, per (1)—among the 15 states participating in the program. Our analysis of the 100,140 records flagged in Iowa in 2012 allows us to better understand how many of these pairings represented duplicate registrations and how many of these duplicate registrations actually produced double votes. Of the 34,900 pairings in which SSN4 is known for both records, 25,987 had the same SSN4. This suggests that roughly three-quarters of the registrations flagged by Crosscheck were in fact duplicate registrations, although election administrators often lack the necessary SSN4 to determine whether a particular match is good or bad. More importantly, fewer than 10 of the known 25,987 duplicate registrations were used to cast a ballot twice. This suggests that both the number of registration records that share common observable characteristics and the number of duplicate registrations are poor proxies for the number of double votes. Reporting the first two quantities in place of the last risks confusing the public about in the integrity of American elections.

Scholars have been concerned about the (mis)measurement of voter fraud because sometimes the twin goals of improving electoral integrity and voter accessibility come into conflict. Many election administration policies fall along a continuum from promoting accessibility, with some potential loss of integrity, to protecting integrity, but potentially disenfranchising legitimate voters. For example, the adoption of absentee ballots made it easier for people to access a ballot, particularly those who are elderly or disabled (Barreto et al., 2006; Miller and Powell, 2016), while also introducing new ways through which fraudulent ballots could be cast (Fund, 2004, p. 47-50). Likewise, when maintaining voter registration records, there is a tradeoff between reducing deadwood and potentially purging legitimate registrations.

One reason that people disagree about how to run elections is that they focus on either accessibility or integrity, without much consideration of this tradeoff. For example, when speaking out against a South Carolina voter identification law, a Democratic state representative argued that "if even one person is disenfranchised because of this law, that will be one person too many" (People For The American Way, 2012). Kris Kobach used similar logic but instead contended that "one case of voter fraud is [one] too many" (Lowry, 2015). Such statements promote a debate that focuses on maximizing accessibility or integrity, without any consideration for the other dimension.

This paper suggests that both Democrats and Republicans would be wise to acknowledge that there are tradeoffs between accessibility and integrity when administering elections. Voter registration list maintenance is necessary, particularly in the United States, where a decentralized election apparatus produces duplicate registrations as people move across jurisdictions. Crosscheck recommends election administrators remove duplicate voters registration records by removing the earlier record. But election administrators must be mindful of the efficacy of any purging strategy, including the number of double votes prevented relative to the legitimate votes impeded. In the case of Crosscheck's proposed purging strategy, this means knowing how many times the duplicate registration with the earlier registration date is used to cast a second, illegitimate, vote, relative to the number of times it is used to cast a single, legitimate vote. Based on the data from Iowa, we find this proposal would impede 300 legal votes for every double vote prevented. This highlights the fact that many policies that would reduce the potential for fraud also make it more difficult for some legitimate votes to be cast. Likewise, many policies that make voting more accessible also increase opportunities for fraud. Emphasizing accessibility or integrity, without consideration for the other, is likely to lead to poor election administration.

References

Ahlquist, John S, Mayer, Kenneth R, and Jackman, Simon (2014). "Alien abduction and voter impersonation in the 2012 US General Election: Evidence from a survey list experiment." *Election Law Journal* 13(4): 460–475.

Alvarez, R Michael, Hall, Thad E, and Hyde, Susan D (2009). "Studying Election Fraud." In: *Election fraud: detecting and deterring electoral manipulation*. Ed. by Alvarez, R Michael, Hall, Thad E, and Hyde, Susan D. Washington, DC: Brookings Institution Press: 1–17.

Ansolabehere, Stephen and Hersh, Eitan (2010). "The quality of voter registration records: A state-by-state analysis." *Institute for Quantitative Social Science and Caltech/MIT Voting Technology Project Working Paper*. URL: http://hdl.handle.net/1902.1/18550.

Ansolabehere, Stephen and Persily, Nathaniel (2008). "Vote fraud in the eye of the beholder: The role of public opinion in the challenge to voter identification requirements." *Harvard Law Review* 121(7): 1737–1774.

Barreto, Matt A. et al. (2006). "Do absentee voters differ from polling place voters? New evidence from California." *Public Opinion Quarterly* 70(2): 224–234.

Cantú, Francisco and Saiegh, Sebastián M (2011). "Fraudulent democracy? An analysis of Argentina's infamous decade using supervised machine learning." *Political Analysis* 19(4): 409–433.

Christensen, Ray and Schultz, Thomas J (2013). "Identifying election fraud using orphan and low propensity voters." *American Politics Research* 42(2): 311–337.

Cox, Gary W and Kousser, J Morgan (1981). "Turnout and rural corruption: New York as a test case." *American Journal of Political Science*: 646–663.

Deckert, Joseph, Myagkov, Mikhail, and Ordeshook, Peter C (2011). "Benford's Law and the detection of election fraud." *Political Analysis* 19(3): 245–268.

Enamorado, Ted, Fifield, Benjamin, and Imai, Kosuke (2017). "Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records." *Working Paper*. URL: http://imai.princeton.edu/research/files/linkage.pdf.

Fund, John (2004). *Stealing elections: How voter fraud threatens our democracy*. San Francisco, CA: Encounter Books.

Hasen, Richard L (2012). *The voting wars*. New Haven, CT: Yale University Press.

Hood, MV and Gillespie, William (2012). "They just do not vote like they used to: A methodology to empirically assess election fraud." *Social Science Quarterly* 93(1): 76–94.

Hopkins, Daniel J. et al. (2017). "Voting but for the law: Evidence from Virginia on photo identification requirements." *Journal of Empirical Legal Studies* 14 (1).

Johnson, Andrew (2014). "N.C. state board finds more than 35K Incidents of "Double Voting" in 2012." *National Review* April 2.

Levitt, Justin (2007). *The truth about voter fraud*. New York, NY: Brennan Center for Justice.

Lowry, Bryan (2015). "Kobach's voter prosecutions draw scrutiny to proof-of-citzenship requirement." *Wichita Eagle* October 18.

Mayer, Jane (2012). "The voter-fraud myth." *The New Yorker* October 29.

McDonald, Michael P. (2007). "The true electorate: A cross-validation of voter registration files and election survey demographics." *The Public Opinion Quarterly* 71(4): 588–602.

McDonald, Michael P and Levitt, Justin (2008). "Seeing double voting: An extension of the birthday problem." *Election Law Journal* 7(2): 111–122.

Mebane, Walter R. (2009). "Election forensics: The second-digit Benford's law test and recent American presidential elections." In: *Election fraud: detecting and deterring electoral manipulation*. Ed. by Alvarez, R Michael, Hall, Thad E, and Hyde, Susan D. Washington, DC: Brookings Institution Press.

Mebane, Walter R (2011). "Comment on Benford's Law and the detection of election fraud?" *Political Analysis* 19(3): 269–272.

Miller, Peter and Powell, Sierra (2016). "Overcoming voting obstacles: The use of convenience voting by voters with disabilities." *American Politics Research* 44(1): 28–55.

Minnite, Lorraine Carol (2010). *The myth of voter fraud*. Ithaca, NY: Cornell University Press.

Montgomery, Jacob M et al. (2015). "An informed forensics approach to detecting vote irregularities." *Political Analysis* 23(4): 488–505.

Morris, Dick (2014). "Investigate 2012 vote fraud." *The Hill* April 8.

People For The American Way (2012). "SC African American Ministers: Voter ID Decision Shows Continued Need for Voting Rights Act." Press Release.

Pew (2012). *Inaccurate, costly, and inefficient: Evidence that America's voter registration system needs an upgrade*. Pew Charitable Trusts.

Sadinle, Mauricio (2017). "Bayesian Estimation of Bipartite Matchings for Record Linkage." *Journal of the American Statistical Association*.

Steorts, Rebecca C., Hall, Rob, and Fienberg, Stephen E. (2016). "A Bayesian Approach to Graphical Record Linkage and Deduplication." *Journal of the American Statistical Association* 111: 1660–1672.

Stewart III, Charles, Ansolabehere, Stephen, and Persily, Nathaniel (2016). "Revisiting public opinion on voter identification and voter fraud in an era of increasing partisan polarization." *Stanford Law Review* 68(6): 1455–1489.

Wand, Jonathan N et al. (2001). "The butterfly did it: The aberrant vote for Buchanan in Palm Beach County, Florida." *American Political Science Review* 55(2): 793–810.

Weigel, David (2016). "For Trump, a new "rigged" system: The election itself." *Washington Post* August 2.

Yancey, William E (2010). "Expected number of random duplications within or between lists." *Proceedings of the Section on Survey Research Methods, American Statistical Association* 2010: 2938–2946.

A Appendix

A.1 Modeling the birthday distribution

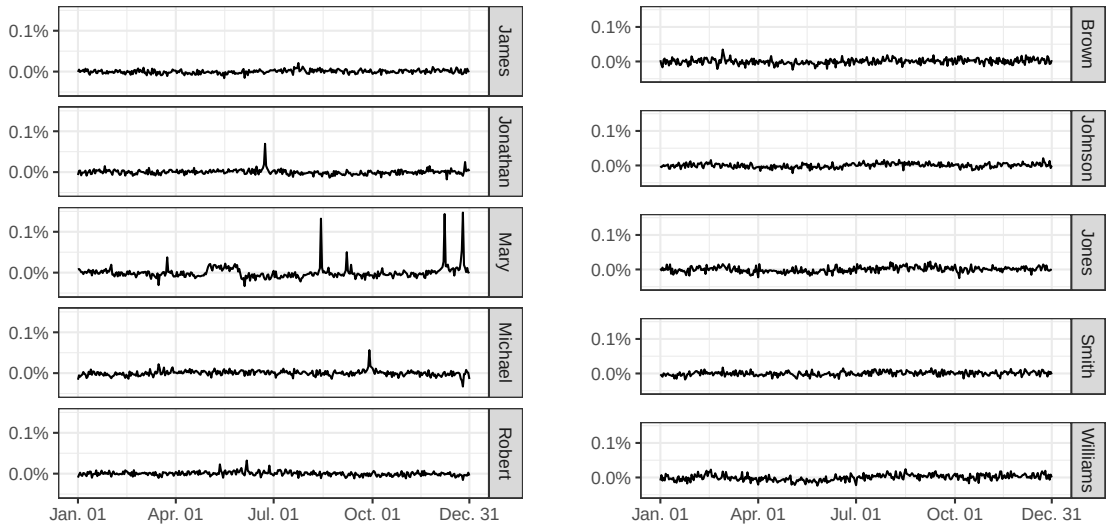
Our goal is to estimate $\Pr(B = b \mid F = f, L = l, Y = y)$, the probability that a voter has a birthday b conditional on having first name f, last name l, and being born in year y. The challenge is that we do not observe a sufficient number of people with the same name who were born in the same year to estimate this only using the empirical distribution. Our first simplification is to assume that $\Pr(B = b \mid F = f, L = l, Y = y) = \Pr(B = b \mid F = f, Y = y)$, so that we can ignore an individual's last name when estimating this probability. The justification for this assumption comes from Figure A.1, which plots the difference in the share of voters with the most common first and last names born on a given day and the share of the general population of voters born on that same day. The left panel of the plot shows a disproportionate number of voters named John and Mary are born on St. John's Day (June 24) and near Christmas, respectively. The right panel does not show similar spikes in the common last names. This pattern is understandable since first names are actively selected whereas last names are generally not. Lemma A.1 derives our estimate of $\Pr(B = b \mid F = f, Y = y)$ under three assumptions.

Lemma A.1 *Assume:*

1. *If* $d_{b,y_1} = d_{b,y_2} \forall b$, *then* $\Pr(B = b \mid Y = y_1, F = f) = \Pr(B = b \mid Y = y_2, F = f)$;
2. $\Pr(F = f, D = d \mid B = b) = \Pr(F = f \mid B = b) \Pr(D = d \mid B = b)$;
3. $\Pr(D = d \mid B = b) = \Pr(D = d)$.

Then we have,

$$\Pr(B = b \mid F = f, Y = y) = \frac{Pr(B = b \mid F = f)\ \Pr(D = d_{b,y})}{\sum_{b'} \Pr(B = b' \mid F = f)\ \Pr(D = d_{b',y})}\ . \tag{10}$$



(a) (b)

Figure A.1: Examples of names among 2012 voters with a non-uniform date of birth distribution, by day (a) or year (b) of birth.

The first assumption means that if y_1 and y_2 are two different years with the same weekday schedule, then the distribution of birthdays for a given first name is the same. Two years have the same weekday schedule when January 1st falls on the same day of the week in both years, and neither or both years are a leap year. Note that while this assumption means that someone named Connor born in 1973 would have the same probability of being born on January 1st as someone named Connor born in 1979, as both were Mondays, it does not require the number of Connors born in 1973 and 1979 to be the same. We use the notation $y' \sim y$ to indicate that year y' has the same weekday schedule as year y.

The second assumption means that the distribution of first names of people born on a given day is independent of the day of the week. So once we condition on being born on a given day, nothing is learned about what day of the week one was born on from one's first name. While we acknowledge there are cases — like being named Wednesday — where this assumption is not correct, such cases are rare.

The third assumption is that birthday and birth day-of-week are independent. Thus, knowing an individual's birthday does not give us any information on the day of the week

they were born on.

A.1.1 Proof of Lemma A.1

Consider the set of people born with first name f and birthday b on day of the week $d_{b,y}$, which is represented by $\{B = b, D = d_{b,y}, F = f\}$. Without loss of generality, we can decompose this set into the union of sets of people born with first name f and birthday b in a year y' such that $d_{b,y'} = d_{b,y}$. Going one step further, and ignoring leap years, we can say that $d_{b,y'} = d_{b,y}$ is equivalent to y' and y having the same weekday schedule, which we can write as $y' \sim y$ using our notation:

$$\{B = b, D = d_{b,y}, F = f\} = \bigcup_{(y' \text{ s.t. } y' \sim y)} \{B = b, Y = y', F = f\}.$$

Because the sets on the right-hand side of the equation above correspond to different years, and thus have no intersection, we can write,

$$\begin{aligned}
\Pr(B = b, D = d_{b,y}, F = f) &= \sum_{(y' \text{ s.t. } y' \sim y)} \Pr(B = b, Y = y', F = f), \\
\Pr(B = b, D = d_{b,y} \mid F = f)\ \Pr(F = f) &= \sum_{(y' \text{ s.t. } y' \sim y)} \Pr(B = b, Y = y' \mid F = f)\ \Pr(F = f), \\
\Pr(B = b, D = d_{b,y} \mid F = f) &= \sum_{(y' \text{ s.t. } y' \sim y)} \Pr(B = b, Y = y' \mid F = f) \\
&= \sum_{(y' \text{ s.t. } y' \sim y)} \Pr(B = b \mid Y = y', F = f)\ \Pr(Y = y' \mid F = f).
\end{aligned}$$

Assumption 1 gives us that $\forall y' \sim y, \Pr(B = b \mid Y = y', F = f) = \Pr(B = b \mid Y = y, F = f))$, so that,

$$\begin{aligned}
\Pr(B = b, D = d_{b,y} \mid F = f) &= \sum_{(y' \text{ s.t. } y' \sim y)} \Pr(B = b \mid Y = y', F = f)\ \Pr(Y = y' \mid F = f) \\
&= \Pr(B = b \mid Y = y, F = f) \sum_{(y' \text{ s.t. } y' \sim y)} \Pr(Y = y' \mid F = f).
\end{aligned}$$

Rearranging terms, we get,

$$\Pr(B = b \mid Y = y, F = f) = \frac{\Pr(B = b, D = d_{b,y} \mid F = f)}{\sum_{(y' \text{ s.t. } y' \sim y)} \Pr(Y = y' \mid F = f)}. \tag{11}$$

Using Bayes' rule, we can rewrite the numerator in Eq. (11) as,

$$\begin{aligned} \Pr(B = b, D = d_{b,y} \mid F = f) &= \frac{\Pr(F = f, D = d_{b,y} \mid B = b)\ \Pr(B = b)}{\Pr(F = f)} \\ &= \frac{\Pr(F = f \mid B = b)\ \Pr(D = d_{b,y} \mid B = b)\ \Pr(B = b)}{\Pr(F = f)} \end{aligned} \tag{12}$$

where the second equality comes from assumption 2, which gives us that $\Pr(F = f, D = d \mid B = b) = \Pr(F = f \mid B = b) \Pr(D = d \mid B = b)$. By Bayes' rule,

$$\Pr(F = f \mid B = b) = \frac{\Pr(B = b \mid F = f)\ \Pr(F = f)}{\Pr(B = b)}. \tag{13}$$

Plugging Eq. (13) into Eq. (12) and simplifying gives us that

$$\begin{aligned} \Pr(B = b, D = d_{b,y} \mid F = f) &= \Pr(F = f \mid B = b) \times \Pr(D = d_{b,y} \mid B = b) \times \frac{\Pr(B = b)}{\Pr(F = f)} \\ &= \frac{\Pr(B = b \mid F = f)\ \Pr(F = f)}{\Pr(B = b)} \times \Pr(D = d_{b,y} \mid B = b) \times \frac{\Pr(B = b)}{\Pr(F = f)} \\ &= \Pr(B = b \mid F = f)\ \Pr(D = d_{b,y} \mid B = b) \\ &= \Pr(B = b \mid F = f)\ \Pr(D = d_{b,y}) \end{aligned} \tag{14}$$

where the final equality comes from assumption 3, which gives us that $\Pr(D = d \mid B = b) = \Pr(D = d)$. Substituting the results of Eq. (14) into the numerator of Eq. (11) gives us that

$$\begin{aligned}\Pr(B = b \mid Y = y, F = f) &= \frac{\Pr(B = b, D = d_{b,y} \mid F = f)}{\sum_{(y' \text{ s.t. } y' \sim y)} \Pr(Y = y' \mid F = f)} \\ &= \frac{\Pr(B = b \mid F = f)\ \Pr(D = d_{b,y})}{\sum_{(y' \text{ s.t. } y' \sim y)} \Pr(Y = y' \mid F = f)} \\ &= \frac{\Pr(B = b \mid F = f)\ \Pr(D = d_{b,y})}{Z(f, y)}. \end{aligned} \tag{15}$$

To solve for $Z(f, y)$ we note that it must be the case that $\sum_{b'} \Pr(B = b' \mid Y = y, F = f) = 1$ for it to be a valid probability distribution. Thus,

$$Z(f, y) = \sum_{b'} \Pr(B = b' \mid F = f)\ \Pr(D = d_{b',y}). \tag{16}$$

Plugging in Eq. (16) to Eq. (15) yields the lemma.

A.2 Proof of Theorem 4.1

To simplify the notation, we represent $M_{f,l,y}$ by M, $N_{f,l,y}$ by N, $D_{f,l,y}$ by D, $p_{b_s|f,l,y}$ by p_s, and $k_{f,l,y}$ by k. We start by computing the expectation of M. For $1 \leq i < j \leq m + k$, let $A_{i,j}$ indicate whether $B_i = B_j$. Then by the linearity of expectation,

$$\mathbb{E}M = \mathbb{E}\left(\sum_{1 \leq i < j \leq m+k} A_{i,j}\right) = \sum_{1 \leq i < j \leq m+k} \mathbb{E}A_{i,j}. \tag{17}$$

For $1 \leq i \leq k$, $\mathbb{E}A_{i,m+i} = 1$ since $B_i = B_{m+i}$ by construction. For the remaining $\binom{m+k}{2} - k$ terms, $\mathbb{E}A_{i,j} = \Pr_D(B_i = B_j) = \sum_s p_s^2$. Consequently,

$$\begin{aligned}\mathbb{E}M &= k + \left(\binom{m+k}{2} - k\right) \sum_s p_s^2 \\ &= k\left(1 - \sum_s p_s^2\right) + \binom{m+k}{2} \sum_s p_s^2.\end{aligned}$$

By rearranging terms, we now have that $\mathbb{E}\hat{k} = k$.

To compute the variance of $\hat{k}$, we first compute the variance of M, decomposing it as

$$\mathrm{Var}(M) = \sum_{1 \le i < j \le m+k} \mathrm{Var}(A_{i,j}) + 2 \sum_{S} \mathrm{Cov}(A_{i,j}, A_{k,l}) \tag{18}$$

where S is the set of indices so that each distinct, unordered pair $(A_{i,j}, A_{k,l})$ appears in the sum exactly once. Since $A_{i,j}$ is an indicator variable,

$$\mathrm{Var}(A_{i,j}) = \mathbb{E}A_{i,j} - (\mathbb{E}A_{i,j})^2 . \tag{19}$$

By the above, $\mathrm{Var}(A_{i,m+i}) = 0$ for $1 \le i \le k$; and for the remaining terms, $\mathrm{Var}(A_{i,j}) = \sum_s p_s^2 - (\sum_s p_s^2)^2$. Consequently,

$$\sum_{1 \le i < j \le m+k} \mathrm{Var}(A_{i,j}) = \left(\binom{m+k}{2} - k \right) \left(\sum_s p_s^2 - \left(\sum_s p_s^2 \right)^2 \right) . \tag{20}$$

Next we consider the covariance terms $\mathrm{Cov}(A_{i,j}, A_{k,l})$, dividing them into two sets and analyzing them separately.

Case 1: We first consider the terms where the indices i, j, k, l are all distinct. If neither B_i nor B_j are copies of either B_k or B_l, then $A_{i,j}$ and $A_{k,l}$ are clearly independent, and so $\mathrm{Cov}(A_{i,j}, A_{k,l}) = 0$. Now suppose that exactly one (but not both) of $\{B_i, B_j\}$ is a copy of either B_k or B_l. In this case, since each observation can be a copy of at most one other observation, B_i cannot be a copy of B_j, and B_k cannot be a copy of B_l. We thus have,

$$\mathbb{E}A_{i,j} = \mathbb{E}A_{k,l} = \sum_s p_s^2 \qquad \text{and} \qquad \mathbb{E}A_{i,j}A_{k,l} = \sum_s p_s^3 .$$

Consequently,

$$\mathrm{Cov}(A_{i,j}, A_{k,l}) = \sum_s p_s^3 - \left(\sum_s p_s^2 \right)^2 .$$

Moreover, there are $2k\left[\binom{m+k-2}{2} - (k-1)\right]$ such instances where there is a single copy between $\{B_i, B_j\}$ and $\{B_k, B_l\}$. To see this, note that we can enumerate the instances by first selecting one of the k copies (and its pair); then selecting two additional observations from the remaining $m+k-2$ while avoiding the $k-1$ combinations that result in selecting another copy and its pair; and lastly, choosing one of the two ways in which the selected observations can be combined to form two unordered pairs.

Finally, suppose that both B_i and B_j are copies of B_k and B_l. As above, B_i cannot be a copy of B_j, and B_k cannot be a copy of B_l, so

$$\mathbb{E}A_{i,j} = \mathbb{E}A_{k,l} = \sum_s p_s^2 \qquad \text{and} \qquad \mathbb{E}A_{i,j}A_{k,l} = \sum_s p_s^2.$$

Consequently,

$$\mathrm{Cov}(A_{i,j}, A_{k,l}) = \sum_s p_s^2 - \left(\sum_s p_s^2\right)^2.$$

There are $2\binom{k}{2}$ such terms, since we must first select two of the k copies, and then select one of the two ways in which to combine the four random variables into two unordered pairs.

Case 2: We next consider the covariance terms where there are three distinct indices among the set $\{i, j, k, l\}$. Since $i \neq j$ and $k \neq l$, this means that $\{i, j\} \cap \{k, l\} \neq \emptyset$. If there are no copies among the three distinct random variables, then

$$\mathbb{E}A_{i,j} = \mathbb{E}A_{k,l} = \sum_s p_s^2 \qquad \text{and} \qquad \mathbb{E}A_{i,j}A_{k,l} = \sum_s p_s^3$$

and so,

$$\mathrm{Cov}(A_{i,j}, A_{k,l}) = \sum_s p_s^3 - \left(\sum_s p_s^2\right)^2.$$

The number of such terms—with three distinct random variables, none of which are copies of one another—is $3\left[\binom{m+k}{3} - k(m+k-2)\right]$. To count the terms, we first count the $\binom{m+k}{3}$ ways of selecting three variables from the $m+k$, and then subtract the number of possibilities in which one variable is a copy of another. This latter quantity can be obtained by first

selecting one of the k copied variables and its pair, and then selecting a third observation from the remaining $m+k-2$. Finally, given the three random variables, we form two pairs by selecting which one of the three to duplicate, and replicating that selected variable in each pair.

Now, if B_i is a copy of B_j, then $A_{i,j} = 1$. Consequently, $A_{i,j}$ and $A_{k,l}$ are independent, and so $\mathrm{Cov}(A_{i,j}, A_{k,l}) = 0$. An analogous argument holds if B_k is a copy of B_l.

Finally, if the non-repeated variable among $\{B_i, B_j\}$ is a copy of the non-repeated variable among $\{B_k, B_l\}$, then

$$\mathbb{E}A_{i,j} = \mathbb{E}A_{k,l} = \sum_s p_s^2 \qquad \text{and} \qquad \mathbb{E}A_{i,j}A_{k,l} = \sum_s p_s^2$$

and so,

$$\mathrm{Cov}(A_{i,j}, A_{k,l}) = \sum_s p_s^2 - \left(\sum_s p_s^2\right)^2.$$

Such terms number $k(m+k-2)$, since we must select a copied random variable and its pair, and then a third random variable among the remaining $m+k-2$ to replicate.

Aggregating all the above terms, we have,

$$\begin{aligned}\mathrm{Var}(M) &= \left[\sum_s p_s^2 - \left(\sum_s p_s^2\right)^2\right]\left[\binom{m+k}{2} - k + 4\binom{k}{2} + 2k(m+k-2)\right] \\ &\quad + \left[\sum_s p_s^3 - \left(\sum_s p_s^2\right)^2\right]\left[4k\binom{m+k-2}{2} - 4k(k-1) + 6\binom{m+k}{3} - 6k(m+k-2)\right].\end{aligned}$$

Since $\mathrm{Var}(\hat{k}) = \mathrm{Var}(M)/\left(1-\sum_s p_s^2\right)^2$,

$$\begin{aligned}\mathrm{Var}(\hat{k}) &= \left[\frac{\sum_s p_s^2}{1-\sum_s p_s^2}\right]\left[\binom{m+k}{2} + 4\binom{k}{2} + 2k(m+k-2) - k\right] \\ &\quad + \left[\frac{\sum_s p_s^3 - \left(\sum_s p_s^2\right)^2}{\left(1-\sum_s p_s^2\right)^2}\right]\left[4k\binom{m+k-2}{2} + 6\binom{m+k}{3} - 4k(k-1) - 6k(m+k-2)\right].\end{aligned}$$

Finally, to derive an upper bound on $\mathrm{Var}(\hat{k})$ that is independent of k, observe that $\sum_s p_s^2 \leq \sum_s p_s = 1$, and so $\sum_s p_s^2/(1-\sum_s p_s^2) \geq 0$. Moreover, by Jensen's inequality applied to the convex function $\phi(x) = x^2$ and weights p_i, $\sum_s p_s^3 \geq (\sum_s p_s^2)^2$. Thus, the two terms involving p_i in the variance expression above are non-negative. Consequently, dropping the negative terms, and noting that $k \leq (m+k)/2$, we get the bound

$$\mathrm{Var}(\hat{k}) \;\leq\; 4\binom{m+k}{2}\left[\frac{\sum_s p_s^2}{1-\sum_s p_s^2}\right] + 12\binom{m+k}{3}\left[\frac{\sum_s p_s^3 - (\sum_s p_s^2)^2}{(1-\sum_s p_s^2)^2}\right].$$

On the other hand, to derive a lower bound, we can minimize positive terms and maximize negative terms in the variance expression. Considering $k \leq (m+k)/2$, observe that $4\binom{k}{2} + 2k(m+k-2) - k \geq -\frac{m+k}{2}$, and $4k\binom{m+k-2}{2} - 4k(k-1) - 6k(m+k-2) \geq -4(\frac{m+k}{2})(\frac{m+k}{2} - 1) - 6(\frac{m+k}{2})(m+k-2) = -4(m+k)(m+k-2)$. So we can write

$$\begin{aligned}\mathrm{Var}(\hat{k}) \;\geq\; & \left[\binom{m+k}{2} - \frac{m+k}{2}\right]\left[\frac{\sum_s p_s^2}{1-\sum_s p_s^2}\right] \\ & + \left[6\binom{m+k}{3} - 4(m+k)(m+k-2)\right]\left[\frac{\sum_s p_s^3 - (\sum_s p_s^2)^2}{(1-\sum_s p_s^2)^2}\right].\end{aligned}$$

A.3 Proof of Lemma 4.1

To simplify the notation, we represent $K_{f,l,y}$ by K, $N_{f,l,y}$ by N, and $k^{\mathrm{orig}}_{f,l,y}$ by k^{orig}. We start by computing the expectation of K. By definition, K is the number of unique objects with a copy observed in the updated set. Initially and before updating the set, there are $n - k^{\mathrm{orig}}$ unique objects out of which k^{orig} objects have a copy in the set, and the remaining $n - 2k^{\mathrm{orig}}$ objects are with no duplicates. Each of these k^{orig} objects will still have a copy in the updated set if and only if neither itself nor its copy is dropped. The probability that an object and its copy are not dropped is $(1-p_r)^2$. For the remaining $n - 2k^{\mathrm{orig}}$ unique objects, each will have copy in the updated set if and only if it gets duplicated, which has a

probability of p_u. Therefore,

$$\mathbb{E}K = k^{\text{orig}}(1-p_r)^2 + (n-2k^{\text{orig}})p_u = k^{\text{orig}}\left[(1-p_r)^2 - 2p_u\right] + np_u. \tag{21}$$

Rearranging terms, we get,

$$\mathbb{E}\left[\frac{K-np_u}{(1-p_r)^2-2p_u}\right] = k^{\text{orig}}. \tag{22}$$

n is the number of objects in the original set, while N is the size of updated set. Each object in the original set contributes two objects to the updated set with probability p_u, or one object with probability $(1-p_u)(1-p_r) = 1-p_u-p_r+p_up_r$. Therefore,

$$\mathbb{E}N = \sum_{i=1}^{n} 2p_u + 1 - p_u - p_r + p_up_r = n(1+p_u-p_r+p_up_r) \tag{23}$$

Substituting $n = \frac{\mathbb{E}N}{1+p_u-p_r+p_up_r}$ into the Eq. (22), we have $\mathbb{E}\hat{k}^{\text{orig}} = k^{\text{orig}}$.

Note that in the proof of Theorem 4.1 we were estimating the number of pairs of duplicates in the set, while here we are interested in the number of unique records with duplicates in the set. As long as we assume a person does not vote more than twice in the election, the two estimation approaches yield the same result.

A.4 Proof of Theorem 4.2

To simplify the notation, we represent $T_{f,l,y}$ by T, $D_{f,l,y}$ by D, $p_{b_s|f,l,y}$ by p_s, and $c_{f,l,y}$ by c. Let us first define $\mathcal{Q}$ to be the set of pairs (B_i, B_j) where $1 \le i < j \le m+k$, B_i and B_j belong to different states, and exactly one of them has its binary voting flag set to one. In other words

$$\mathcal{Q} = \{\ (B_i, B_j)\ \mid\ 1 \le i < j \le m+k\ \ ,\ \ 1 \le \nexists u \le l : \{B_i, B_j\} \subset S_u\ \ ,\ \ f_i \oplus f_j = 1\ \}.$$

Here, $f_i \oplus f_j = 1$ means exactly one of f_i and f_j is set to one.

Based on this notation, T is the number of pairs $(B_i, B_j) \in \mathcal{Q}$ such that $B_i = B_j$, and c is the number of cases for $1 \leq i \leq k$ where $(B_i, B_{m+i}) \in \mathcal{Q}$.

Let $A_{i,j}$ indicate whether $B_i = B_j$. Then by the linearity of expectation,

$$\mathbb{E}T = \mathbb{E}\left(\sum_{(B_i,B_j)\in\mathcal{Q}} A_{i,j}\right) = \sum_{(B_i,B_j)\in\mathcal{Q}} \mathbb{E}A_{i,j}. \tag{24}$$

For all the (B_i, B_j) pairs in $\mathcal{Q}$ for which $j = m+i$, $B_i = B_j$ by construction, so $\mathbb{E}A_{i,j} = 1$. By definition, the number of these pairs is c. For the remaining $|\mathcal{Q}| - c$ pairs, $\mathbb{E}A_{i,j} = \Pr_D(B_i = B_j) = \sum_s p_s^2$. Consequently,

$$\begin{aligned}\mathbb{E}T &= c + (|\mathcal{Q}| - c)\sum_s p_s^2 \\ &= c\left(1 - \sum_s p_s^2\right) + |\mathcal{Q}|\sum_s p_s^2.\end{aligned}$$

To compute $|\mathcal{Q}|$, we first count all the (B_i, B_j) pairs where $i < j$ and exactly one of f_i and f_j is set to one. This count is equal to number of ways we can choose a pair with first element from observations with flag set to one ($\sum_{l=1}^{u} v_i$ observations) and second element from observations with flag set to zero ($\sum_{l=1}^{u} \bar{v}_i$ observations), which sums up to $\sum_{l=1}^{u} v_l \sum_{l=1}^{u} \bar{v}_l$. Then we eliminate the pairs where B_i and B_j are from the same set. For each set S_l, we need to eliminate $v_l\bar{v}_l$ such pairs. Therefore,

$$|\mathcal{Q}| = \sum_{l=1}^{u} v_l \sum_{l=1}^{u} \bar{v}_l - \sum_{l=1}^{u} v_l\bar{v}_l.$$

By substituting $|\mathcal{Q}|$ and rearranging terms, we now have that $\mathbb{E}\hat{c} = c$.

To compute the variance of $\hat{c}$, we first decompose variance of T as

$$\mathrm{Var}(T) = \sum_{(B_i,B_j)\in\mathcal{Q}} \mathrm{Var}(A_{i,j}) + 2\sum_S \mathrm{Cov}(A_{i,j}, A_{k,l}) \tag{25}$$

where S is the set of (i, j, k, l) indices such that each distinct unordered pair from elements in $\mathcal{Q}$ appears in the sum exactly once. For $A_{i,j}$ we can write,

$$\mathrm{Var}(A_{i,j}) = \mathbb{E}A_{i,j} - (\mathbb{E}A_{i,j})^2 . \tag{26}$$

For all the (B_i, B_j) pairs in $\mathcal{Q}$ for which $j = m + i$, $\mathbb{E}A_{i,j} = 1$. Therefore, for those pairs $\mathrm{Var}(A_{i,j}) = 0$. There are c such pairs in $\mathcal{Q}$, and for the remaining $|\mathcal{Q}| - c$ pairs, $\mathrm{Var}(A_{i,j}) = \sum_s p_s^2 - (\sum_s p_s^2)^2$. Consequently,

$$\sum_{(B_i,B_j)\in\mathcal{Q}} \mathrm{Var}(A_{i,j}) = (|\mathcal{Q}| - c)\left(\sum_s p_s^2 - \left(\sum_s p_s^2\right)^2\right). \tag{27}$$

Next we consider the covariance terms $\mathrm{Cov}(A_{i,j}, A_{k,l})$. By Cauchy-Schwarz's inequality,

$$\mathrm{Cov}(A_{i,j}, A_{k,l}) \le \sqrt{\mathrm{Var}(A_{i,j})\mathrm{Var}(A_{k,l})}. \tag{28}$$

If either (B_i, B_j) or (B_k, B_l) are among the c pairs in $\mathcal{Q}$ for which one observation is a copy of another, then $\mathrm{Var}(A_{i,j})\mathrm{Var}(A_{k,l}) = 0$. For all the other cases, $\mathrm{Var}(A_{i,j})\mathrm{Var}(A_{k,l}) = \left(\sum_s p_s^2 - (\sum_s p_s^2)^2\right)^2$. Therefore,

$$\sum_S \mathrm{Cov}(A_{i,j}, A_{k,l}) \le \binom{|\mathcal{Q}| - c}{2}\left(\sum_s p_s^2 - \left(\sum_s p_s^2\right)^2\right). \tag{29}$$

Combining equations for terms in $\mathrm{Var}(T)$, we can write,

$$\mathrm{Var}(T) \le (|\mathcal{Q}| - c)^2\left(\sum_s p_s^2 - \left(\sum_s p_s^2\right)^2\right). \tag{30}$$

Consequently,

$$\begin{aligned}\mathrm{Var}(\hat{c}) &= \mathrm{Var}(T)/\left(1-\sum_s p_s^2\right)^2 \\ &\leq (|\mathcal{Q}|-c)^2\left[\frac{\sum_s p_s^2}{1-\sum_s p_s^2}\right].\end{aligned}$$

To make the bound on $\mathrm{Var}(\hat{c})$ independent of c, we substitute $|\mathcal{Q}|-c$ by $|\mathcal{Q}|$ and replace it with the previously calculated count, which yields to

$$\mathrm{Var}(\hat{c}) \leq \left(\sum_{l=1}^{u} v_l \sum_{l=1}^{u} \bar{v}_l - \sum_{l=1}^{u} v_l \bar{v}_l\right)^2 \left[\frac{\sum_s p_s^2}{1-\sum_s p_s^2}\right].$$

A.5 Simulation

We evaluate the performance of our estimation strategy on synthetic datasets with a known number of double votes and which preserve key features of the real data, including correlations between names and dates of birth. To create each synthetic dataset, we carry out the following procedure, starting with the preferred version of the voter file.

1. Randomly select a year-of-birth and first name pair from the voter file.
2. Randomly, and independently of Step 1, select a last name from the voter file.
3. Given the selected first name, last name, and year of birth triple, generate a birthdate based on the modeled birthdate distribution $\hat{p}_{b|f,l,y}$.
4. Repeat the above three steps until the size of the sample equals the size of the voter file.
5. Randomly select k vote records in the synthetic dataset and add copies of them to the synthetic dataset.

This procedure preserves the correlation between first names and dates of birth, including year. By randomly and independently selecting last names, we add additional variance to

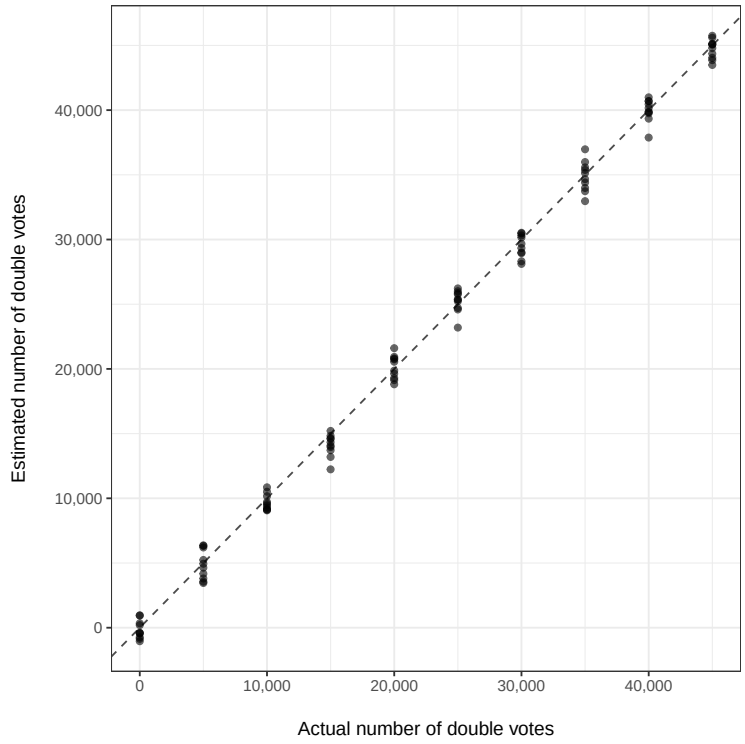


Figure A.2: Estimated number of duplicate records in a simulation compared to actual number of records duplicated.

the dataset. Before duplicating any records, all observed matches are purely coincidental, and thus the full synthetic dataset has exactly k true double votes.

On each synthetic dataset, we carry out our full double vote estimation procedure, including fitting a model to estimate the distribution of $p_{b|f,l,y}$. Figure A.2 shows the result of this exercise on 100 synthetic datasets generated as above for a range of values for k. We find that our estimates are very well aligned with the true number of double votes in these datasets.

We use an analogous procedure to generate bootstrap estimates of variance for our empirical double vote estimate. Specifically, we generate 100 synthetic datasets as above, with k equal to our double vote point estimate, and then compute the variance of our 100 estimates on the synthetic datasets. This procedure can be viewed as a parametric bootstrap, as we use our estimated birthday model to generate the bootstrap samples.

A.6 Measurement Error in Voter File

To estimate the number of people who voted twice in the 2012 election, we use Target Smart's national voter file, which lists the first name, middle name,[17] last name, suffix, date of birth, and turnout history associated with a voter registration.[18] These data provide a nearly comprehensive list of 2012 general election participation: the data include 124,942,823 vote records from the 2012 election, as compared to the 129,085,410 votes cast for a presidential candidate nationwide.[19] Before using the data, we standardize first names in the voter file by converting nicknames to their canonical form. We use pdNickname software, which contains tables relating nicknames to canonical names. We only consider short form or diminutive nicknames with the highest relationship quality scores (less than 5). If a nickname maps to multiple canonical names, we convert it to the most popular canonical name among voters with the same gender. For instance, a male voter named Chris is considered Christopher, and a female voter named Chris is considered Christine.

One concern with these data is that date of birth may not always be reported accurately in the voter file. Figure A.3 shows the distribution of birthdays (i.e., month and day of birth) for voter registrations with a birth year of 1970 and a vote record in 2012. It illustrates a pattern, also shown by Ansolabehere and Hersh (2010), that too many registration records indicate that a voter was born on first day of the month. Across all years, about 14% of 2012 vote records are indicated to have been born on the first day of the month.[20] Such

[17]The data also include middle name, although we do not use this information in our analysis. First, states do not require middle name to be reported and not everyone has a middle name. Among those who both have a middle name and report it, the information is often recorded inconsistently. Many records also contain only a middle initial, making it difficult to assess the accuracy of a given match. Other records have what appear to be transcription errors, such as a suffix in the middle name field.

[18]Some states do not reveal the full date of birth on each registration. In such cases, Target Smart supplements the missing birthdates with information obtained from commercial data sources.

[19]http://www.fec.gov/pubrec/fe2012/federalelections2012.pdf

[20]We can detect some other seemingly random clumps of birthdays in a few states. For instance, March 26th in Wisconsin and New Hampshire, June 5th in Idaho, and the whole month of January in Hawaii all

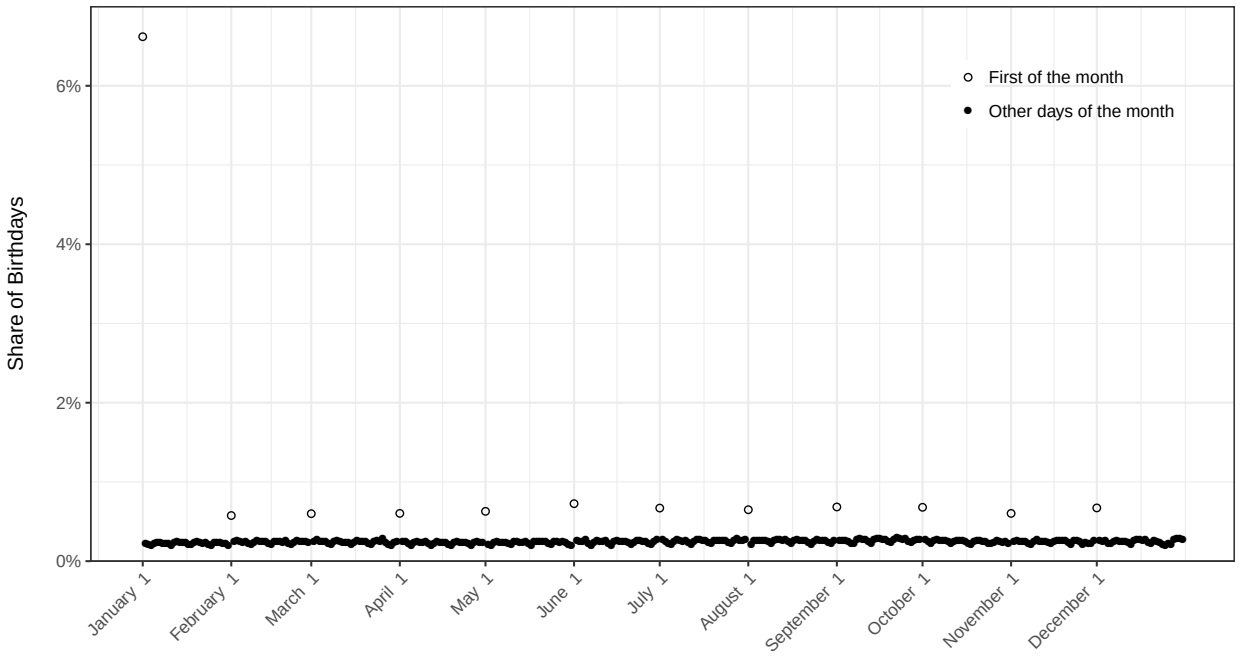


Figure A.3: Distribution of birthdays in 1970 in the voter file.

measurement error could cause us to incorrectly count two votes cast by distinct voters as instead coming from a single voter, and thus overestimate the true rate of double voting.

We also suspect that the birthdates of individuals in multi-generational households are reported incorrectly in a few states. When we match vote records within states by not only first name, last name, and date of birth, but also registration address, we find 7,504 and 2,350 in-state duplicate voters in Mississippi and Wisconsin, respectively. In a vast majority of these cases, the records share a different middle name or suffix, suggesting a situation in which either a father (mother) or son (daughter) were assigned the others' birthdate. Figure A.4 shows the distribution of potential multi-generational matches within states, normalized based on the size of the state. In addition to Wisconsin and Mississippi, we see that the District of Columbia, Arkansas, New Hampshire, Hawaii, and Wyoming also have a disproportionate number of cases in which voter records with the same observable characteristics reside in the same household. These issues in multi-generational households

show a higher concentration of certain voter registration birthdays

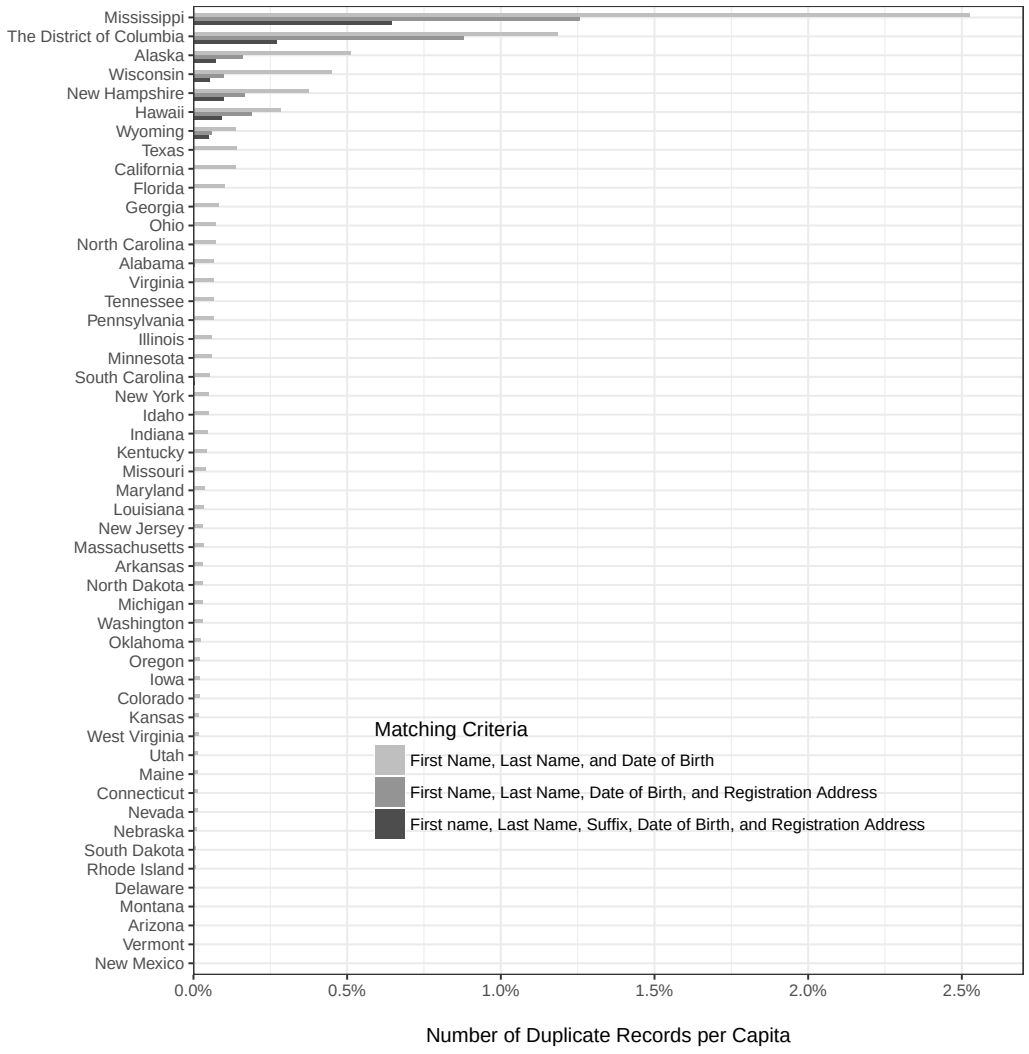


Figure A.4: Distribution of potential multi-generational matches within a state.

raise broader concerns about the quality of the voter file records in these states. We thus exclude these states from our preferred sample, and then scale-up our estimates to account for their removal when generating our final, national numbers.

Finally, we carry out a simulation to assess the sensitivity of our results to possible birthdate errors that may remain in our preferred sample. Given an error rate p, we randomly select $p\%$ of records in our preferred sample and assign each a new birthdate chosen uniformly at random from days in the recorded birth year. We then estimate the number of double votes in the synthetic dataset by running it through our full analysis pipeline, including

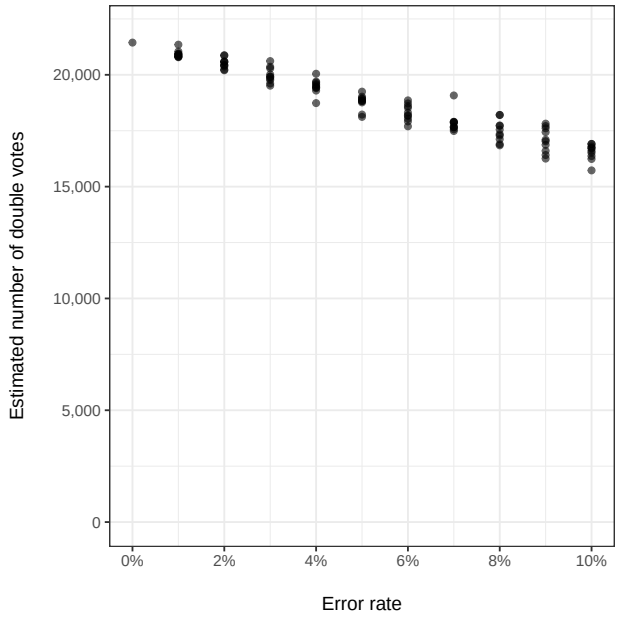


Figure A.5: Estimated number of double votes in the preferred sample from multiple simulations as we increase the error in recording of birthdates.

estimation of $p_{b|f,l,y}$. Figure A.5 shows the result of this procedure when we simulate 10 synthetic datasets for each error rate p in the range 1% to 10%. We see that an error rate of p corresponds to an approximately $2p$ reduction in the estimated number of double votes. To understand why, note that any actual case of double voting in our synthetic datasets becomes undetectable with probability approximately equal to $2p$, since each vote record in the pair has probability p of being assigned a new birthdate. This explanation, however, only holds approximately, as birthdate errors also attenuate the day-of-week effect, among other factors, complicating theoretical analysis and prompting our simulation.

A.7 Estimating Measurement Error in Vote Records

Ansolabehere and Hersh (2010) present the best evidence constructed to date on the accuracy of vote records in voter files. For each county in a given election, Ansolabehere and Hersh calculate the absolute value of the deviation between number of vote records in the voter file minus the total number of ballots cast in the certified aggregate returns. They

aggregate these deviations over all of the counties in the state and divide by the total number of votes cast in the state. From this analysis, Ansolabehere and Hersh conclude that about two percent of voter registrations are incorrectly classified as having voted or abstained.

There are two primary limitations of this analysis. First, Ansolabehere and Hersh's method does not allow us to distinguish between false negatives and false positives, leaving open the possibility that there are few false positives. Second, their method also would understate the amount of measurement error in counties in which some registrations are wrongly classified as abstaining, while others are wrongly classified as voting.

We use the data collected from our Philadelphia poll book audit to estimate the rate at which registrations not used to vote are incorrectly given an electronic vote record (i.e., a false positive). There were 17,586 electronic registration records that did not have an electronic record of voting in these precincts.[21] In 33 of these cases, we found the registration had a record of being used in the poll book. We also found 144 cases in which a registration was listed as voting in the electronic records, but had no record of having voted in the poll book (i.e., a signature discrepancy) and 29 cases of a registration being listed as voting in the electronic records, but not being listed in the poll book (i.e., a registration discrepancy). This suggests that the $f_p = \frac{144+29}{17{,}586+144+29-33} = 0.0098$.

Of course, we cannot be certain that these records are all false positives. It could be the case that the electronic voting records are correct and the poll book fails to note it. One way to indirectly assess this possibility is to compare the rates at which voter registrations with signature and registration discrepancies were recorded as voting in the elections leading up to 2010. If the previous vote history of these registrants is similar to the previous vote history of registrants who did not vote in 2010, this would suggest that many of these records are false positives. Conversely, if the previous vote history of these registrants is similar to the previous vote history of registrants who did vote in 2010, this would suggest that registrants with signature and registration discrepancies represent errors in the poll book, and thus are

[21] A few additional records couldn't be validated because pages were missing in the poll books.

Table A.1: Examining Past Vote History of 2010 Signature and Registration Errors

	Dependent variable: Electronic record of voting in			
	2006	2007	2008	2009
	(1)	(2)	(3)	(4)
2010 electronic voting record	.448	.395	.436	.248
	(.005)	(.005)	(.005)	(.004)
Signature discrepancy	-.174	-.159	-.084	-.135
	(.042)	(.039)	(.033)	(.029)
Registration discrepancy	-.361	-.396	-.123	-.189
	(.079)	(.048)	(.076)	(.048)
Potential false negative	.224	.250	.357	.133
	(.085)	(.082)	(.067)	(.062)
Constant	.170	.083	.461	.018
	(.003)	(.002)	(.004)	(.001)

Note: N = 29,263 registered voters in the 47 precincts that were audited.

not false positives.

Table A.1 suggests that some, but not all, of the signature and registration discrepancies are false positives. To benchmark the past turnout of those who did and did not vote in 2010, we first calculate the 2006 turnout rate of those we know to have voted and not voted in 2010. Table A.1 shows that 62% of 2010 voters also turned out in 2006, while only 17% of those who abstained in 2010 participated in 2006. The 2006 turnout behavior of those with signature or registration discrepancies in 2010 falls somewhere in between, at 44% and 26%, respectively. We see similar patterns for 2007, 2008, and 2009 turnout as well. The fact that those with discrepancies between the electronic records and poll books previously voted at a rate somewhere in between those who abstained and those who voted in 2010 suggests that the false positive rate is both greater than zero and less than 1.0%.

These audit results are meant only to be illustrative, not representative, of the false positive rate in the population. There are some reasons why the false positive rate in Philadelphia may be larger than the rate in the general population. Ansolabehere and Hersh (2010) found that there were more discrepancies than average in Pennsylvania between the

number of ballots cast and the number of vote records in the voter file. And while a majority of jurisdictions either used Philadelphia's poll-book-and-bar-code approach or a voter sign-in sheet with no bar codes, a small, but growing number of jurisdictions, use an electronic poll book, particularly in states with early voting.[22] Because electronic poll books remove the step in which poll books are translated into electronic records, use of such technology is likely to reduce the number of false positives.

However, there are also reasons why we might expect there to be fewer false positives in Philadelphia than in the general population. Because of the size of the jurisdiction, the Philadelphia Voter Registration Office has a large, professionalized, and experienced staff that it can draw upon when scanning the poll books. And while there is more potential for error using the poll-book-and-bar-code approach than using electronic poll books, even more error is likely to occur in places that manually key-in the information contained in the poll book. It is also the case that there are false positives that our audit would not detect. For example, a poll worker could sign in a voter under the wrong registration. Consistent with this, Hopkins et al. (2017) report that 105 individuals had to resort to filing a provisional ballot in Virginia during the 2014 midterm election after they arrived at their polling place to find their registration was wrongly marked as having been used to vote earlier in the day.

Because we only have a rough sense of the rate of false positives, it is hard to say anything definitive about how many of the potential double votes can be explained by measurement error. Ultimately, all we can conclude is that measurement error likely explains a sizable portion, and possibly nearly all, of the surplus double votes that we observe in the national voter file.

[22]The Election Administration and Voting Survey suggests about 15% and 25% of voters used such technology in 2008 in 2012, respectively.

Table A.2: Robustness Checks on 2012 Analysis in Table 1

Target Smart (TS) or Vote File (VF)						
to Measure Iowa Turnout	TS		VF		VF	
Drop States with > 10%						
First of Month Birthdays	No		No		Yes	
SSN4 Match	Yes	No	Yes	No	Yes	No
Which Reg. Used to Vote:						
Both	7	1476	8	1489	1	420
One (earlier reg. date)	2543	1687	2695	1756	700	632
One (later or unknown reg. date)	9429	2572	9882	2649	2040	431
Neither	14008	3178	13402	3019	5285	934

A.8 Measurement Error Linking Vote Records to Crosscheck Data

Table A.2 suggests that measurement error in turnout does not affect our conclusion that few likely double votes were identified in the Crosscheck data. Columns 3 and 4 replicate our 2012 analysis when Iowa turnout is linked to the Crosscheck data from the voter file using Iowa's voter registration number. While we find one additional case of a likely double vote, we also find more than a hundred additional cases in which only the Iowa registration was used to cast a vote. We expand upon this analysis in columns 5 and 6 by limiting the sample of states paired to Iowa to those states in which fewer than 10% of 2012 voters have a birthday on the first of the month. We do this because we expect there to be fewer cases in which we fail to match a vote record to a registration record in these states. We find a single likely double vote out of the 421 potential double votes in these states. Moreover, we find 700 cases in which only the voter registration record with the earlier registration date was used to cast a ballot.

	Grid of Potential Duplicate Voters Within States by DOB Last Name First Name														
2012	AZ	AR	CO	IL	IA	KS	KY	LA	MI	MS	MO	NE	OK	SD	TN
AZ		2,829	24,863	16,014	7,153	3,687	688	2,062	27,617	2,220	7,569	3,306	4,006	2,449	3,614
AR	2,829		4,557	6,950	2,430	2,686	691	5,957	5,085	6,477	11,049	995	7,403	433	7,180
CO	24,863	4,557		19,902	10,850	10,035	1,054	5,065	17,086	3,309	12,498	8,927	8,306	3,937	6,153
IL	16,014	6,950	19,902		31,882	6,311	2,467	5,207	49,260	10,766	39,658	3,803	4,834	1,500	12,469
IA	7,153	2,430	10,850	31,882		4,706	526	1,558	7,019	1,797	11,563	10,954	2,031	4,865	2,806
KS	3,687	2,686	10,035	6,311	4,706		401	1,369	4,461	1,397	31,082	4,196	6,575	905	2,205
KY	688	691	1,054	2,467	526	401		873	2,267	1,085	1,195	233	576	117	1,905
LA	2,062	5,957	5,065	5,207	1,558	1,369	873		6,851	17,744	5,254	810	2,829	277	4,422
MI	27,617	5,085	17,086	49,260	7,019	4,461	2,267	6,851		7,527	12,960	2,416	4,067	1,265	16,956
MS	2,220	6,477	3,309	10,766	1,797	1,397	1,085	17,744	7,527		5,607	780	2,364	305	21,661
MO	7,569	11,049	12,498	39,658	11,563	31,082	1,195	5,254	12,960	5,607		4,244	7,539	1,300	7,804
NE	3,306	995	8,927	3,803	10,954	4,196	233	810	2,416	780	4,244		1,126	2,608	1,108
OK	4,006	7,403	8,306	4,834	2,031	6,575	576	2,829	4,067	2,364	7,539	1,126		402	2,858
SD	2,449	433	3,937	1,500	4,865	905	117	277	1,265	305	1,300	2,608	402		537
TN	3,614	7,180	6,153	12,469	2,806	2,205	1,905	4,422	16,956	21,661	7,804	1,108	2,858	537	
Totals	108,077	64,722	136,542	211,023	100,140	80,016	14,078	60,278	164,837	83,039	159,322	45,506	54,916	20,900	91,678

Figure A.6: Distribution of potential duplicate voters in 2012 according to internal documents circulated by the Interstate Crosscheck Program.

A.9 Additional Figures

A.10 Additional Tables

Table A.3: Estimated Double Votes in Sample by Sample Restriction and Birthday Distribution

	(1)	(2)	(3)	(4)
Drop First of Month Birthdays	No	Yes	Yes	Yes
Drop States with Multigenerational Issues	No	No	Yes	Yes
Keep Commercially Sourced Birthdays	Yes	Yes	Yes	No
Number of Vote Records in Sample (millions)	124.94	107.62	104.21	102.65
Coverage in Sample of Total Votes (FEC)	.968	.834	.807	.795
Vote Record Pairings with Same First and Last Name and DOB	3047626	826455	761875	737463
Estimated Double Votes in Sample by Distribution:				
Birthday Distribution Uniform	2065273	47252	28694	27674
	(1991)	(1768)	(1715)	(1687)
Birthday Distribution Conditional on Year	155607	44067	25856	24927
	(7792)	(1773)	(1719)	(1691)
Birthday Distribution Conditional on Year and First Name	408001	39394	21441	20619
	(6920)	(1781)	(1727)	(1699)

Table A.4: Robustness Checks on Estimated Double Votes in Population

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Drop States with Multigenerational Issues	Yes	No	Yes	Yes	Yes	Yes	Yes
Standardize First Names	Yes	Yes	No	Yes	Yes	Yes	Yes
Smoothing Parameter in Birthday Distribution (θ)	10200	10200	10200	1000	50000	10200	10200
Keep Observations With Commercial Birthday	Yes	Yes	Yes	Yes	Yes	No	Yes
Analytical (A) or Bootstrapped (B) Standard Errors	A	A	A	A	A	A	B
Estimated Double Votes in Sample	21441	39394	20237	19863	22992	20619	21441
Sample-to-Population Scale Factor	1.534	1.439	1.534	1.534	1.534	1.580	1.534
Estimated Double Votes in Population	32890	56688	31044	30470	35270	32599	32890
	(2649)	(2563)	(2462)	(2652)	(2645)	(2686)	(1405)

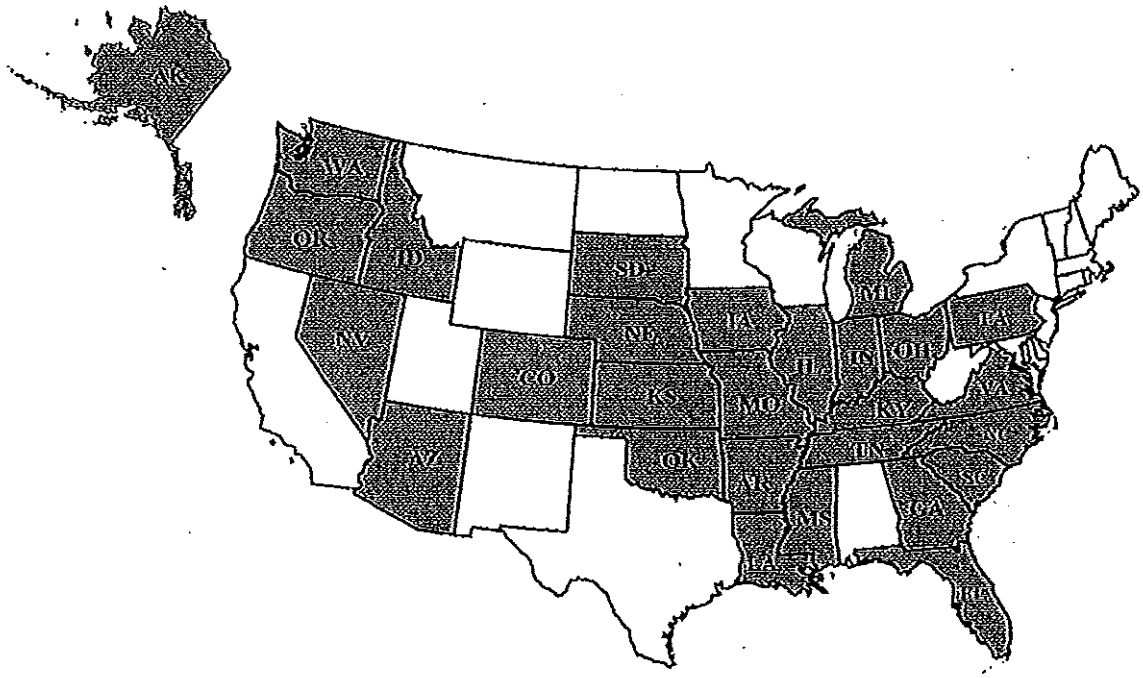
A.11 Crosscheck 2014 Participation Guide

Interstate Voter Registration Data Crosscheck

2014 Participation Guide

December, 2013

Alaska, Arizona, Arkansas, Colorado, Florida, Georgia, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Michigan, Mississippi, Missouri, Nebraska, Nevada, North Carolina, Ohio, Oklahoma, Oregon, Pennsylvania, South Carolina, South Dakota, Tennessee, Virginia and Washington.



Contents

I. Joining the Crosscheck Program

1. Chief State Election Official (CSEO) or designee signs the Memorandum of Understanding (MOU)

2. CSEO assigns two staff members:
    a. one election administration person
    b. one IT/database person

3. Staff members will:
    a. participate in annual conference calls and emails
    b. pull voter registration data in January and upload to FTP site
    c. receive crosscheck results and process them
    d. instruct local election officials
        (1) mail notices to registrants
        (2) promptly respond to requests for signatures, addresses, etc.

4. There is no cost. Processing the duplicate registrations and researching possible double votes requires a commitment of time at the state and local levels. States make individual decisions about the amount of time and effort they will commit, and this might vary from year to year. A state that is not able to commit the resources to process the results in a given year still provides a benefit to the other states through its participation.

II. Data Comparison Procedure

1. Designate at least one administrative and one IT/database contact person to be on the email list serve.

2. At least one person from each state should participate in a conference call hosted by Kansas in November or December preceding the crosscheck.

3. Pull your entire database on January 15, 2014 and upload it to the secure FTP site hosted by Arkansas. Instructions and, if necessary, followup reminders and questions, will come from the Kansas or Arkansas IT/database persons. Each state's data should include Active and Inactive records if possible.
Follow the prescribed data format. Review and edit your data before uploading it. Do not truncate fields, reverse fields, or leave them blank. Please include SSN4.

4. Kansas will download files, delete everything from FTP site, run the data comparison, and upload individual state results files to the FTP site. At every stage of the process, data files are encrypted and zipped.

5. When notified by Kansas, download your results files from the Arkansas FTP site. After downloading, make sure all data are deleted from the FTP site.

6. Process the results according to your state's laws, regulations and policies.

7. Respond promptly to inquiries from other states or local jurisdictions for information to confirm duplicates or to obtain evidence of double votes. Usually this will be copies of signatures on poll books or absentee/advance ballot applications and return envelopes.

III. Analyzing Results

Each state analyzes and acts upon the results according to its own laws and regulations. No state is required by the Memorandum of Understanding to act upon the results.

A. Cancellations and Confirmation Mailings

An apparent duplicate registration is produced when the first names, last names and dates of birth in two records match exactly. Other information such as middle name, suffix and SSN4 should be used to confirm whether the two records are matches. It may be necessary to contact another jurisdiction to obtain more information, such as signatures.

An apparent duplicate registration may result in one of two actions being taken:

1. The jurisdiction possessing the record with the older registration date may cancel the record (and send a cancellation notice if state laws or regulations require it) if the following conditions are met:
    a. The records match on first name, last name, and date of birth, and
    b. One or both of the following data elements match:
        - last four digits of Social Security number and
        - signature and
    c. Data in the middle name field either matches or is not a mismatch.

2. The jurisdiction possessing the record with the older registration date may mail a confirmation notice, pursuant to the National Voter Registration Act of 1993, Sec. 8(d)(2), if the three fields match as specified in item 1.a. above. These registrants' names are added to the state's Inactive list pending cancellation after two federal general elections, assuming there has been no voting activity during that period.

B. Cancellations by Confirmation Between Jurisdictions

Pursuant to NVRA Sec. 8(d)(1)(A), the jurisdiction possessing the record with the older registration data may cancel the record (and send a cancellation notice if state laws or regulations require it) if another jurisdiction confirms that the registrant has registered to vote in the newer jurisdiction and has indicated on the voter registration application form an address in the former jurisdiction.

C. Information Sharing

Each state will decide whether it prefers that followup requests for information from other states and localities should be addressed to the state or the individual

localities. All participating states will be notified of this preference. Each state will provide contact information for local election offices. States and localities are cautioned against sending registrants' personally identifiable information via email.

D. Double Votes

1. When two records are determined to be duplicates, review the voter history field to determine if there appears to be a double vote. Experience in the crosscheck program indicates that a significant number of apparent double votes are false positives and not double votes. Many are the result of errors—voters sign the wrong line in the poll book, election clerks scan the wrong line with a barcode scanner, or there is confusion over father/son voters (Sr. and Jr.).

2. Collect copies of signatures from the election officers in the two jurisdictions in which the double votes occurred. The classic double vote occurs when a person votes in person at the polling place on election day in the jurisdiction where he/she normally lives and also casts an absentee (advance) ballot by mail in the other jurisdiction.
In these cases, evidence to prove the double vote occurred often includes the following:

- Signature from the voter's application for voter registration in jurisdiction A
- Signature from the voter's application for voter registration in jurisdiction B
- Signature from the poll book in jurisdiction A
- Signature on an absentee (advance) ballot application form in jurisdiction B
- Signature on the absentee (advance) ballot return envelope in jurisdiction B

3. The collection of evidence to prove double votes is a considerable commitment of time and effort. It requires a high level of cooperation and communication between jurisdictions.

4. Compare the signatures. Once you are satisfied that the evidence indicates a double vote occurred, refer the case to a local or state prosecutor. Include a referral cover letter, cite relevant state statutes, and include copies of all necessary documents.

E. Information Request Form

A request form as appears below may be used to request followup information from other jurisdictions. Jurisdictions may adapt it as needed and produce it on their own letterhead.

KRIS W. KOBACH
Secretary of State

Memorial Hall, 1st Floor
120 S.W. 10th Avenue
Topeka, KS 66612-1594
(785) 296-4564

STATE OF KANSAS

12/4/2013

To Whom It May Concern:

Voter Information

John Doe	DOB:	1/02/1933
Jane Doe		2/03/1955
Tom Smith		3/06/1985
Janet Jones		12/13/1967
Ben Thompson		11/9/1990

We request voter registration and voter history information related to the above mentioned individuals for the November 2012 election. The purpose of the request is to collect evidence about possible double votes cast by these individuals.

We will maintain appropriate safeguards to protect the confidentiality of the records.

We will not make any public use of these files or information. We will keep your office apprised of the details as our office moves forward with this inquiry.

If you have questions please contact me at 785-296-0080.

Sincerely,

Jameson Beckner
Special Programs Coordinator
Kansas Secretary of State

IV. IT/Database Manager Information

IT/database managers should follow this timeline and use the data format on the next page.

ACTIVITY	Time Frame
Kansas sends data extract reminder email	January
Arkansas sends upload instruction email to each state with: • URL for FTP site • login ID • password	January
States extract their data according to Data Format document	approx. January 15th
States upload their extract files to the FTP site	January
Each state emails bruce.ferguson@sos.ks.gov • with encryption password • with number of records	January
Kansas processes the extract file	January - February
Kansas emails notification to each state	January - February
Kansas loads the file into comparison database	January - February
Kansas produces Results file for each state: • create Single Row comparison files • create Stacked Row comparison files • update Statistics spreadsheet • zip all comparison files and statistics into Results file • encrypt Results file into self-decrypting .exe • upload Results file to the FTP site	January - February
Kansas sends email to each state that Results file is ready	January - February
Each state needs to: • refer to State Cross Check Result File Instructions below • download their Results file • delete their Results file from the FTP site • decrypt and unzip their Results file	January - February
Each state processes its Results file accordingly	January - February
Kansas and Arkansas verify Results have been deleted from FTP site	February

Data Format

Fields	
1. Status	("A" – Active, "I" – Inactive)
2. DateTime_Generated	
3. First_Name	
4. Middle_Name	
5. Last_Name	
6. Suffix_Name	
7. Date_of_Birth	(YYYY/MM/DD Example: "2010/01/01")
8. Voter_ID_Number	
9. SSN_Last4	
10. Address_Line_1	(if no mailing address, provide residential address)
11. Address_Line_2	
12. City	
13. State	
14. Zip	
15. County_Name	
16. Date_of_Registration	(YYYY/MM/DD Example: "1970/01/01")
17. Voted_in_Last_General	("Y" – they did vote, or "N" – they did not vote, or "" – data not available)

The file should be a comma delimited ASCII file with double quote text qualifiers and {CR} {LF} row delimiters. The file should have a Header Record followed by 1 to many Voter Records. Each Voter Record should contain 17 fields.

Example:

"Status","DateTime_Generated","First_Name","Middle_Name","Last_Name","Suffix_Name", "Date_of_Birth","Voter_ID_Number","SSN_Last4","Address_Line_1","Address_Line_2", "City","State","Zip","County_Name","Date_of_Registration","Voted_in_Last_General"

"A","2013/01/15 12:00:00 AM","Bob","Alan","Jones","","1940/06/16","123456","7890", "123 Main St","Apt 201","Topeka","KS","12345","Shawnee","1958/06/17","Y"

The file should be encrypted and password protected and uploaded to the secure FTP site. Please email the password in a separate email.
Also, please notify us of the total number of records in the uploaded file.

We use a free program, AxCrypt, for encryption. Here is a link to the AxCrypt download site: http://www.axantum.com/AxCrypt/ .

Reminders for Data Upload Process

We have identified from past experience some helpful hints that we ask you to keep in mind as you prepare to upload your data. Please carefully review your file before uploading, taking into account the following:

1. Do not include any records that contain programming commands from your process that created the file
2. Please include a header record, but only one
3. Please account for all 17 fields, in the order requested
4. Please trim all excess spaces so the records are not padded to a fixed length
5. Please zip your file before uploading it to the FTP site
6. Please be aware if your address lines contain a comma and make sure you encapsulate the field with double quotes
7. Please note that if you will be providing SSN data, we only ask for the last 4 positions
8. Please extract dates in the requested formats
9. Please edit data that contains double quotes during extraction – ie. remove the double quotes or change them to single quotes
    a. Examples
        1.. Change …,"Robert "Bob"",…
        To …,"Robert 'Bob"",…

        2. Change …,"123 "U" St",…
        To …,"123 U St",…
10. Consider replacing "null" text values with actual null string, ""

Crosscheck Results File Instructions

1) Download your state's self decrypting .exe file from the FTP site
2) Double click the file.
3) Enter the passphrase
    i) This will decrypt the file
    ii) The resulting .zip file contains 2 folders and a spreadsheet
4) Extract all files
5) Please delete your state's file from the FTP site once you confirm a successful download

- The spreadsheet presents some general statistics about current and previous State Cross Check Voter Registration Comparisons
  - Since DOB is one of the match criteria, please provide valid Date_of_Birth fields (see column D)
  - If you see a non-zero value in column E, please determine if you can provide unique Voter_ID_Number fields
  - If you see a non-zero value in column F, please determine within your own extract file if you have multiple records for the same individual (the criteria for this comparison is the same as the state-to-state comparison; DOB, LastName, FirstName

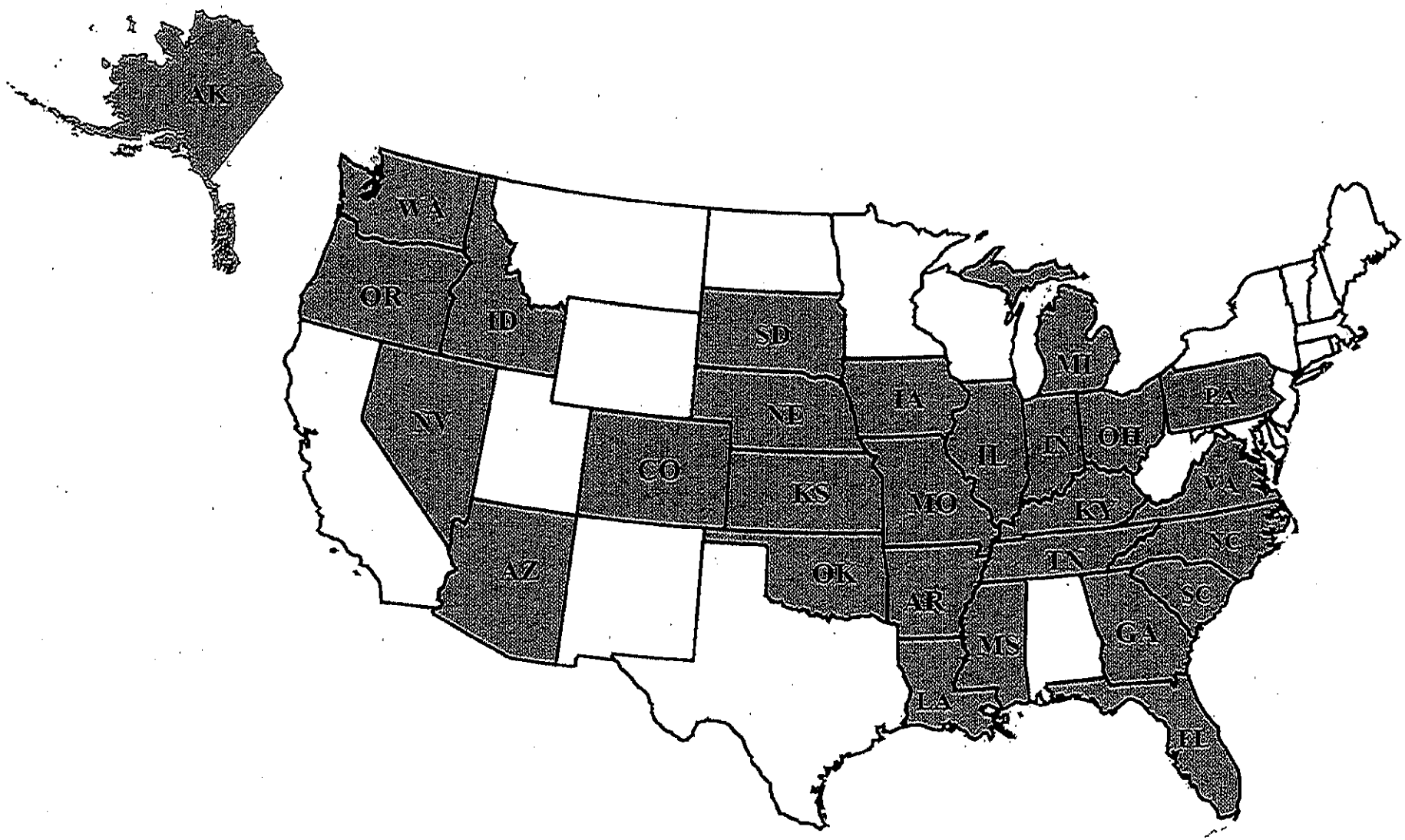
- One folder, SingleRowOutput:
  - Contains one result file with the potential match count of each comparison and total for that BaseState
  - Contains individual result files for your state compared with each other participating state
    - Ideally opened programmatically
  - Contains one result file comparing your state with all other participating states
    - Ideally opened programmatically
  - Within each result file:
    - A header row identifies each column
    - A possible voter match is presented in a single row with your states' data followed by the data from the other state

- The second folder, StackedRowOutput:
  - Contains one result file with the potential match count of each comparison and total for that BaseState
  - Contains individual result files for your state compared with each other participating state
    - Ideally opened in Excel
  - Contains one result file comparing your state with all other participating states
    - Ideally opened in Excel
  - Within each result file:

- A header row identifies each column
- A possible voter match is presented in two rows with data from your state stacked over data from the other state

➢ Within each individual result file:
- The "Case" column represents the sequential instance of each possible match

➢ Within the ALL result file:
- The "Case" column represents the sequential instance of each possible match as that match relates in the individual result files



2014 Interstate Crosscheck