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

January 13, 2017

Abstract

There are more than two 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 voter files are a completely accurate account of who voted, we estimate about 0.02% of the votes cast in 2012 were double votes. An audit of poll books, however, suggests that many of these apparent double votes represent measurement error when recording turnout in voter files. Nevertheless, concerns about double voting have led many states to participate in the Interstate Crosscheck Program, which promotes purging registration records that share a common name and date of birth. We find their proposed purging strategy would eliminate about 200 registrations used to cast legitimate votes for every one registration used to cast a double vote.

Keywords: voting, election law, voter suppression, voter fraud

*We thank Target Smart for supplying us with a national voter file and audience members at the Yale Behavioral Sciences Workshop for their comments and suggestions.

1 Introduction

Election administration is a contentious political issue, with Democrats and Republicans increasingly embroiled in what has been popularly termed the "voting wars" (Hasen, 2012). At the heart of this partisan debate lies different beliefs about the incidence of voter fraud (Ansolabehere and Persily, 2008; Stewart III, Ansolabehere, and Persily, 2016). Such fraud could come in many forms, such as voter impersonation, non-citizen voting, or double voting, the latter of which occurs when the same individual casts multiple ballots, each under a different registration record, in the same election. There have been few successful prosecutions for voter fraud, even in states that have aggressively pursued the issue (Noble, 2013; Lowry, 2015). This type of evidence is an example of what Hood and Gillespie (2012) term "secondary data," which also includes election fraud allegations, election incident reports, and journalistic accounts. Based on these data, Levitt (2007) and Minnite (2010) conclude there has been little to no voter fraud in recent U.S. elections. Not everyone, however, accepts these conclusions. Voter fraud is difficult to detect, and even more challenging to prove beyond a reasonable doubt, when it is done well (Ahlquist, Mayer, and Jackman, 2014). Moreover, the qualitative evidence brought to bear by Levitt and Minnite, while impressive in its breadth, 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 new methodologies are developed to quantify and clarify how often it occurs (Alvarez, Hall, and Hyde, 2009). To this end, the growth of election forensics has ushered in a host of new measurement methods to detect patterns consistent with various conceptions of electoral fraud (see e.g. Mebane, 2009; Fukumoto and Horiuchi, 2011; Beber and Scacco, 2012; Christensen and Schultz, 2013; Montgomery et al., 2015). In this paper, 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 quantify the number of double voters in this election using the information contained in a national voter file.

Double voting occurs when the same individual votes multiple times in the same election. In order to vote in the United States, a person must first 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. Vote records include the metadata which is part of a registration record. When we mention the personal characteristics of a particular vote record, we are referring to the personal characteristics associated with the registration record that was marked as having been used to cast a ballot. While each registration can only be associated with one vote, the same individual may cast more than one vote, for example, by registering to vote, and casting a ballot, in multiple states. This illustrates an important distinction between a voter and a vote: 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 using only official records is challenging in the United States because, unlike in many other countries, there is no national voter registration number. Rather, the only identifiers consistently reported in vote records are first names, last names, and dates of birth. While we observe that 2,892,574 pairs of 2012 vote records in a national voter file share the same first name, last name and date of birth, some of these cases represent two distinct voters who share a common first name, last name, and date of birth rather than a double voter.

Roughly, 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 date of birth from the 2,892,574 observed matching pairs. To estimate the number of distinct voters that we expect to share the same first name, last name, and date of birth, we expand upon a probabilistic birthdate model first developed in McDonald and Levitt (2008). While McDonald and Levitt assume that birthdays are distributed uniformly within a year, our model accounts for the fact that people are more likely be born on certain days. We also extend their study by producing analytic confidence intervals, explicitly accounting for the possibility that some non-voters are marked as voting by error, and looking at the entire country instead of a single state.

Our results show that the amount of double voting is substantially less than Morris' million. If we assume that electronic records are a completely accurate account of who voted and scale our estimate to account for voter records that are missing from our analysis, we estimate that around 30,000 people (0.02% of voters) voted twice in the 2012 presidential election. However, if roughly one percent of registrations were erroneously marked as being used to vote, we would expect to estimate there were around 30,000 double voters even if there were no actual double votes cast.

Errors when translating poll books to electronic records are one reason why a registration could be incorrectly marked as being used to cast a ballot. To get a sense of the rate of translation error, we conducted an audit where we hand coded a set of poll books in Philadelphia, PA to compare the vote records in the primary source poll books to the constructed vote records in the electronic voter file. This audit suggested that there is sufficient measurement error linking poll books to voter files to explain some, and possibly all, of the excess double votes we found in the national voter file.

These statistical estimates dispel some of the more extreme allegations of rampant voter fraud, but Democrats and Republicans may still be stuck in a stalemate. This reflects, in part, an ideological difference. While Democrats focus on the accessibility of the ballot, arguing that "if even one person is disenfranchised... that will be one person too many," Republicans emphasize electoral integrity, countering that "one case of voter fraud is [one] too many."[1] Such statements fail to acknowledge the tradeoffs inherent in policies that increase

[1] See, for example, here for the former and here for the latter.

accessibility or preserve integrity. While deadwood registrations increase the possibility of double voting, purging voter registrations will also increase the chance of "impeded votes," a term we introduce to describe the purging of registration records that would be used to cast a legal vote.

We conclude the paper by evaluating the implicit tradeoff made by the Interstate Crosscheck Program, which compared individual-level voter registration data across 15 and 28 states in 2012 and 2014, respectively, to aid in the purging of duplicate records and prosecution of double voters. States that participate in Crosscheck submit their voter registration data to a centralized repository, and receive back a list of cases in which a registration record with the same first name, last name, and date of birth was identified in another state. We obtained the list of potential duplicate registrations that Crosscheck returned to Iowa prior to the 2012 presidential and 2014 midterm elections, and matched it to a national voter file in order to assess the number of double voters that Crosscheck identified.

Our analysis of Crosscheck data makes explicit the large tradeoff in accessibility sacrificed in an administrative environment that only focuses on maintaining electoral integrity. While Crosscheck flagged more than 5,000 cases across the two elections in which a vote record from Iowa and a vote record from another state shared the same first name, last name, and date of birth, nearly all of these matches featured vote records with inconsistent middle names and different social security numbers (SSN4), and thus were likely two distinct individuals' vote records. In fewer than ten of these cases did both flagged vote records share consistent middle names and the same SSN4s. In contrast, there were thousands of cases in which two registrations shared consistent middle names and the same SSN4s, but only one registration had a record of voting. While Crosscheck advises that in such circumstances the registration record with the earlier registration date be purged, there were over a thousand cases in which this earlier registration record was actually the one used to cast the single vote.

2 The Measurement of Voter Fraud

Voter-driven fraud has received significant media attention as part of what Hasen (2012) terms the "voting wars." Examples of voter-driven fraud, or voter fraud for short, include the casting of multiple ballots (double voting), illegal ballots (e.g., non-citizen voting), or other people's ballots (voter impersonation). This form of electoral fraud should be distinguished from institutional or administrative fraud, such as the fraudulent counting of votes, voter suppression, and bribery. One well-known example of the latter is when President Lyndon Johnson eked out his first election to the Senate with the help of election administrators in Jim Wells County, Texas, who changed Johnson's vote total in the county from 765 to 965, netting him a pivotal 200 votes (Caro, 1990). While this form of fraud is part of political lore, it has been much less relevant to the current policy debate over proposed election laws, like photo identification and proof of citizenship requirements, that are principally about protecting the integrity of ballots, as opposed to the integrity of how these ballots are counted.

One motivation for the wave of such laws is the belief by a substantial portion of the mass public that voter fraud has corrupted the integrity of the electoral process.[2] About 10% of all respondents to a 2008 national survey reported that double-voting and non-citizen voting (considered together) as well as voter impersonation (considered separately) is "very common," while about another 30% felt it "occurs occasionally" (Ansolabehere and Persily, 2008). Suspicion of voter fraud is particularly strong among self-identified Republicans, who were about three times more likely than Democrats to believe double-voting was very common and about twice as likely to feel the same about voter impersonation. Stewart III, Ansolabehere, and Persily (2016) show that these large partisan differences have continued through the present day.

While partisanship colors the view of many facets of American life (Bartels, 2002), the lack of consensus documented above is particularly problematic because of the importance

[2] Others assert that it is about partisan (Hicks et al., 2015) or racial (Bentele and O'Brien, 2013) politics.

of the public perceiving the electoral process as legitimate. 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, is it particularly important for this to be documented and for the public to be made aware.

The most prominent attempts to measure the amount of voter fraud – both of which were undertaken in the last ten years – have dismissed allegations of rampant voter cheating of any stripe (Levitt, 2007; Minnite, 2010). Surveying a variety of what Hood and Gillespie (2012) term "secondary data," Levitt concludes that "voter fraud is extraordinarily rare," while Minnite charges that "voter fraud is a politically constructed myth." One typical piece of secondary evidence brought to bear is the lack of criminal prosecutions for voter fraud. For example, Matt Schultz, the Republican secretary of state of Iowa, secured only five guilty pleas for voter fraud after an 18-month investigation (Noble, 2013).[3] Another piece of evidence common to both Levitt and Minnite are administrative records. For example, Minnite investigates hundreds of allegations made to the California Election Fraud Investigations Unit between 1994 and 2006 and finds that, of the more than 75 million votes cast in statewide elections during this period, only 46 were found to constitute a violation, of which 37 lacked intent. Levitt also details, and debunks, allegations made in newspaper articles about double voting.

These qualitative findings, however, have fueled — rather than quieted — speculative allegations of voter fraud along partisan lines. Such qualitative data are ideal for a court of law, but fail to address the anxieties of the mass public. The type of 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, but which incidents are covered in

[3] Three of the five people convicted were ex-felons, who may have been confused about their voting rights because of a recent change in the state's disenfranchisement law (Noble, 2013).

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, because, as Christensen and Schultz (2013) point out, perhaps only the worst attempts of election fraud leave behind such evidence.

A growing group of scholars, working under the general umbrella of election forensics, have pioneered new quantitative methods that overcome some of these limitations by shifting the level of analysis. Rather than trying to detect individual cases of voter fraud, these methods are designed to estimate the aggregate amount of fraud in the population at-large. Thus, while these methods cannot address the merits of any specific case of voter fraud, they can estimate the prevalence of fraud more broadly, offering empirical evidence that can be used to inform the mass public's beliefs.

Many of these methods examine aggregated data to detect anomalous election outcomes (see e.g. Cox and Kousser, 1981; Powell, 1989; Baum, 1991; King, 2001). Wand et al. (2001), for example, develop ecological analyses to detail how the butterfly ballot led to George W. Bush's pivotal Florida victory in the 2000 presidential election. In many cases, such anomalies would be consistent with institutional fraud. For example, Mebane (2009) develops a test that looks at whether some 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, Myagkov, and Ordeshook (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) deploy machine learning techniques to study improprieties in international elections.

Other methods focus on detecting voter fraud. Ahlquist, Mayer, and Jackman (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) look for evidence consistent with fraud on statewide voter rolls. Hood and Gillespie match a Georgia registration list to the state death registry and find little evidence of ballots being cast in the name of deceased registrants. Christensen and Schultz (2013) look at the voting behavior of low propensity and orphan voters — defined as those who voted in a low-profile election but skipped the preceding and subsequent high-profile elections — to identify potential cases of voter fraud. Their method — which relies on qualitative investigations of statistical anomalies — confirms known cases of various forms of electoral fraud but finds no additional instances of misbehavior in Ohio, Florida, or Utah.

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 likely pretext for the purging of voter rolls (Levitt, 2007). Political elites have continued to hype the threat of voters casting multiple ballots. The current chairman of the Republican National Committee has argued that voter fraud "is a real and persistent threat to our electoral system" and taken to describing double-voters in particular as one of Democrats' "core constituencies" (Mayer, 2012). The conservative pundit Dick Morris has gone into more detail about suspected double-voting: "You're talking about probably over a million people that voted twice in this election," he's claimed, citing a misleading statement by the Interstate Crosscheck Program as "the first concrete evidence we've ever had of massive voter fraud" (Morris, 2014). More recently, 2016 Republican presidential nominee Donald Trump expressed concern that "we may have people vote 10 times" in a "rigged" election (Weigel, 2016).

It is challenging to identify whether the same person voted in multiple states because of the limited identifiers we observe in voter files. Unlike many other countries, the United States doesn't have national voter registration numbers and the only identifiers consistently reported about voters are their first names, last names, and dates of birth.[4] While it is unlikely that any two randomly selected vote records would share a common first name, last

[4] While information about more discriminating identifiers, such as Social Security numbers, are sometimes collected, they typically are not shared for privacy reasons.

name, and birthdate, a sizable number of these cases will occur once we aggregate over the quadrillions of pairs of vote records in the population. As we discuss in greater detail later, we observe more than two million cases in a national voter file in which 2012 vote records share a common first name, last name, and date of birth. While we could collect auxiliary information for any given pair to try and determine whether it represents a case of the same person voting twice or two distinct voters with the same first name, last name, and date of birth, 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 (2008), who study allegations of rampant double voting in New Jersey in 2004, offer the best work on the topic and a starting point for our work here (see also Yancey, 2010). They consider the birthday problem and theorize an extension — the "birthdate problem" — using simulation. By randomly drawing the year of birth for a vote record from the empirical age distribution of registered voters, and assuming that birthdays within years follow a uniform distribution, McDonald and Levitt determine that we should expect 487 cases in which two vote records within New Jersey share the same first name, last name, and date of birth. They find 761 matched pairs in the state, using a matching method similar to Hood and Gillespie (2012) and Christensen and Schultz (2013), and use a simulation to offer a 95% confidence interval of plus or minus 102 matches. They conclude by noting some limitations of their methods, which makes it difficult to know whether their result should be interpreted as evidence of a small number of double voters or the systematic underestimation of the likelihood that two distinct voters in New Jersey share the same names and date of birth. McDonald and Levitt point to two measurement problems in their approach that might lead to such an underestimate of the expected number of in-state duplicates: they do not take into account either name and date of birth periodicity or election official error.

We agree with their self-assessment. In the sections that follow, we outline our approach to the study of double voting, taking care to address both of these issues of measurement as we expand the scope of analysis using a national voter file.

3 A Policy Tradeoff

Scholars have been concerned about the (mis)measurement of voter fraud because improving electoral integrity has implications for voter accessibility. Scott Walker, the governor of Wisconsin, summarized many people's goals for election administration when he said that "we want it to be easy to vote but hard to cheat."[5] Sometimes, though, these twin goals 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 electoral administrative policy 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."[6] Kris Kobach, the current Republican Secretary of State of Kansas, used similar logic but instead contended that "most people realize that one case of voter fraud is too many" (Lowry, 2015). Scott Walker echoed this in a 2014 gubernatorial debate, arguing

[5] Walker has said this on multiple occasions on Twitter. See here for a 2015 tweet and here for a 2016 one.

[6] http://www.pfaw.org/press-releases/2012/10/african-american-ministers-voter-id-decision-shows-continued-need-voting-righ

that “it doesn’t matter if there’s one, 100 or 1,000” illegal ballots, because “who would like to have [their] vote canceled out by a vote that was cast illegally?” (Hohmann, 2014). Such statements promote a debate that focuses on maximizing accessibility or integrity, without any consideration for the other dimension.

A concern about double voting has led many states to participate in the Interstate Crosscheck Program. What began as a regional cooperative a decade ago rapidly expanded to include over half of states by 2014. The program, led by Kris Kobach, compares individual-level voter registration data across participating states to purge duplicate records and prosecute duplicate voters. According to Crosscheck’s Participation Guide (see section A.9 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 date of birth 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.

To date, no scholarly work investigates the tradeoffs involved in administering a program like Crosscheck. From the perspective of electoral integrity, the benefit of such a program is that it could help identify and purge a registration that would be otherwise used to cast a second ballot. The greater the share of the potential double voters that are actual double voters, the greater the potential benefit in terms of electoral integrity. From the perspective of accessibility, the cost of a program like Crosscheck is that it could be used to purge a registration that would be otherwise used to cast a legitimate ballot. Crosscheck currently reports all pairs of registrations that share the same first name, last name, and date of birth, but live in different states. But there is no systematic evidence about the likelihood that such registrations belong to the same person. The lower this likelihood, the greater the potential threat to accessibility. Even when a county is confident that the same person is registered to vote in multiple states, it is not clear how easy it is to distinguish which registration is currently being used. Crosscheck's 2014 Participation Guide advises that in this situation the registration with the earlier registration date can be purged without noting the frequency with which this is, in fact, the deadwood registration (see section A.9 in the appendix). The more that registrations with the earlier — rather than later — registration date are used to cast a single vote, the greater the potential threat to accessibility.

4 Data

4.1 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, middle name, last name, suffix, date of birth, and turnout history associated with a voter registration.[7] These data provide a nearly comprehensive list of 2012 general election participation. However, there are several data quality issues. A disproportionate number of 2012 vote records are associated with a first-of-the-month birthday and a number of states appear to misreport the birthdates of individuals in multi-generational households. We also are concerned that two registration records, with similar, but not identical, first names, may be associated with the same person. We discuss each of these issues in more detail in Section A.5 in the Appendix.

We minimize the influence of such measurement errors on our estimated rate of double voting, we exclude some vote records when running our main analysis. Specifically, we exclude records indicating that someone was born on the first day of the month and records from seven states with a disproportionate number of records that share the same first name, last name, date of birth, and registration address. To maximize our coverage, we include observations with a date of birth imputed from commercial data sources. Thus, our preferred dataset includes 104,206,281 of the 124,942,824 vote records contained in the full dataset.

[7] 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.

In the results section, we explain how we scale our estimates to account for the vote records that were dropped.

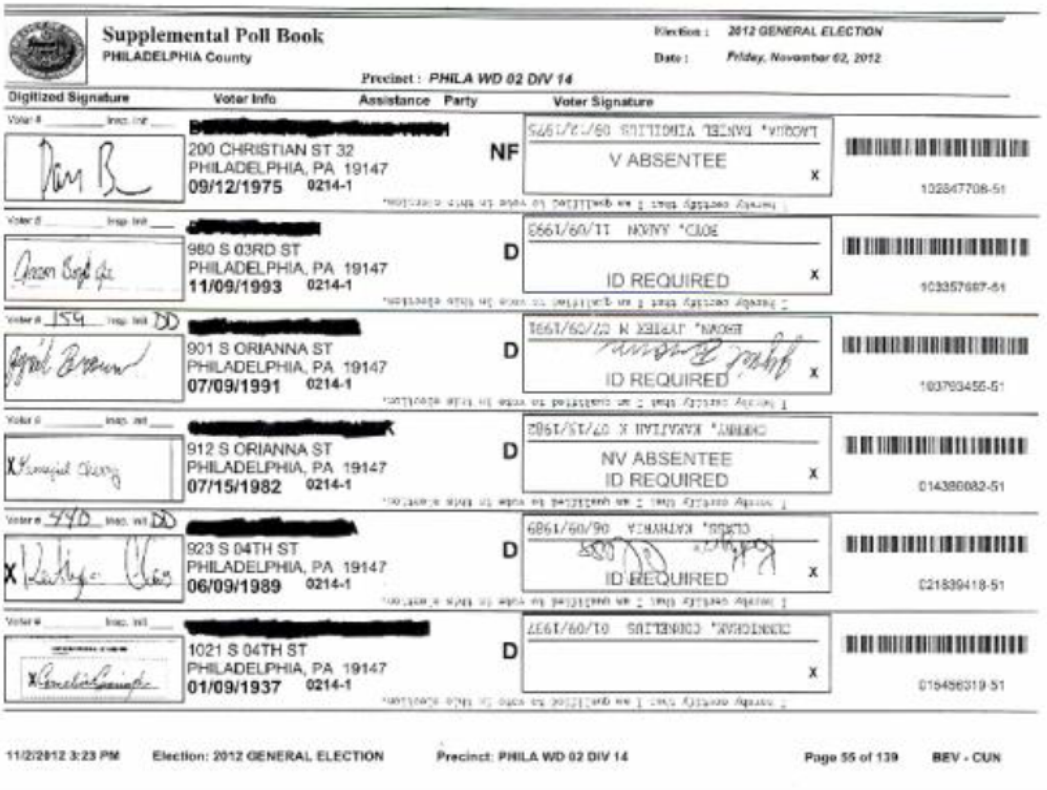
4.2 Philadelphia Voter File Audit

To get a 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.[8] 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. Each record has a bar code. After the election, an employee of the voter registration office scans the bar code next to each registration with a record of voting to generate an electronic record that a vote was cast using that particular registration.

Our principle interest is identifying false positives: registrations with an electronic record of having voted, even though they were not actually used to vote. To identify potential false positives, we searched for registrations that had an electronic record of voting, but were not listed as having voted in a poll book. Auditors validated 11,663 electronic vote records in 47 randomly-selected precincts, documenting two different types of discrepancies.[9] A signature discrepancy refers to the case where a registration record is listed in the voter file for a given precinct, is listed in that precinct's poll book, has an electronic record of voting, but is not listed as having voted in the poll book. A registration discrepancy refers to the case where a registration record is listed in the voter file for a given precinct, has an electronic record of voting, but is not listed in that precinct's poll book.

[8] This voter file was produced on December 8, 2010.

[9] An additional 15 records couldn't be validated because pages were missing in the poll books.



Supplemental Poll Book
PHILADELPHIA County

Election : 2012 GENERAL ELECTION
Date : Friday, November 02, 2012

Precinct : PHILA WD 02 DIV 14

Digitized Signature	Voter Info	Assistance	Party	Voter Signature	
Voter # ____ Insp. Init ____	[redacted] 200 CHRISTIAN ST 32 PHILADELPHIA, PA 19147 09/12/1975 0214-1		NF	V ABSENTEE x	102847708-51
Voter # ____ Insp. Init ____	[redacted] 980 S 03RD ST PHILADELPHIA, PA 19147 11/09/1993 0214-1		D	ID REQUIRED x	103357687-51
Voter # 159 Insp. Init DD	[redacted] 901 S ORIANNA ST PHILADELPHIA, PA 19147 07/09/1991 0214-1		D	ID REQUIRED x	103793456-51
Voter # ____ Insp. Init ____	[redacted] 912 S ORIANNA ST PHILADELPHIA, PA 19147 07/15/1982 0214-1		D	NV ABSENTEE ID REQUIRED x	014386082-51
Voter # 540 Insp. Init DD	[redacted] 923 S 04TH ST PHILADELPHIA, PA 19147 06/09/1989 0214-1		D	ID REQUIRED x	021839418-51
Voter # ____ Insp. Init ____	[redacted] 1021 S 04TH ST PHILADELPHIA, PA 19147 01/09/1937 0214-1		D	x	015456319-51

11/2/2012 3:23 PM Election: 2012 GENERAL ELECTION Precinct: PHILA WD 02 DIV 14 Page 55 of 139 BEV - CUN

Figure 1: Illustration of a Philadelphia poll book.

4.3 Crosscheck

We obtained a copy of the data that the Iowa Secretary of State's Office received from Crosscheck prior to the 2012 and 2014 elections. Iowa received a list of 100,140 and 139,333 voter registrations which shared the same first name, last name, and date of birth as a registration in another state in 2012 and 2014, respectively. For each registration, these data included the middle name, suffix, registration address, county of registration, voter registration number, date of voter registration, and voter registration status (i.e., active or inactive). All of these variables were reported both as listed in the Iowa voter file and the voter file of the state of the matched registration. While the data that Crosscheck sent to Iowa also contained the last four digits of a registration's Social Security number (SSN4), Iowa removed this variable before providing us with these data. Instead, Iowa included an indicator for whether a registration's SSN4 was missing in Iowa, whether the SSN4 was missing in the other state, and whether the SSN4 was the same in Iowa and the other state.

We assess the likelihood that the matches reported by Crosscheck represent the same individual registered twice or two distinct individuals, each registered a single time. Crosscheck reports two pieces of information about both registrations that are useful for making this assessment: SSN4 and the registration's middle name, neither of which are used as part of Crosscheck's matching process. We classify both the SSN4s and middle names within a given pair as being either consistent, inconsistent, or of unknown consistency. When a SSN4 is reported for both registration records, the SSN4s are classified as being consistent if they match exactly, and inconsistent otherwise. If one, or both, of the registration records does not report the SSN4, it classified as being of unknown consistency.

Classifying the consistency of the middle name is slightly more complicated because the same middle name may be presented slightly differently in the two registration records. When two non-empty middle names contain the same number of characters, the middle names are considered consistent if they are exactly the same, and otherwise inconsistent. If the two non-empty middle names contain a different number of characters, let l be the number of characters in the shorter of the two middle names. A middle name is consistent if the first l characters in the longer middle name exactly matches the shorter middle name. If one, or both, of the registration records does not report a middle name, it is classified as being of unknown consistency.

When both the SSN4s and the middle names are consistent, we classify the Crosscheck pair as likely being the same person. Conversely, when either the SSN4s or the middle names are inconsistent, we classify the Crosscheck pair as being unlikely to be the same person. In all other cases, we classify the Crosscheck pair as possibly being the same person.

We also wish to assess the frequency with which votes are cast using the registration records flagged by Crosscheck. We merged the Crosscheck data to the TargetSmart national voter file to learn the vote history associated with registration records. Unfortunately, there were no common identifiers between the Crosscheck and the TargetSmart data. Instead, we exactly matched records on first name, middle name, last name, date of birth, and state. A hand-coding of Crosscheck records that failed to match to the TargetSmart voter file suggests that there are a few people in the Crosscheck data who voted, but failed to match to the TargetSmart voter file, because of discrepancies in the date of birth between the two sources. Thus, we slightly underestimate the share of people listed in the Crosscheck data who voted.

To minimize the number of cases in which we fail to match a voter registration record in Crosscheck to its vote record in the TargetSmart data, we restrict our analysis to states in which TargetSmart registration records appear to have good coverage on date of birth. Specifically, we sum up the total number of vote records in the state in TargetSmart with a birthday not on the first of the month and divided this total by the number of ballots cast in the state according to official records. We limit our analysis to states in which this quantity is greater than 0.9, which reduces our sample size to 64,092 and 97,307 cases in 2012 and 2014, respectively.

5 Results

5.1 Baseline Analysis

We begin by deriving 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. Finally, we consider $p_{b|f,l,y}$: the probability of having a birthday b conditional on having first name f, last name l, and being born in year y.[10] Whereas we can directly observe the first two quantities from the

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

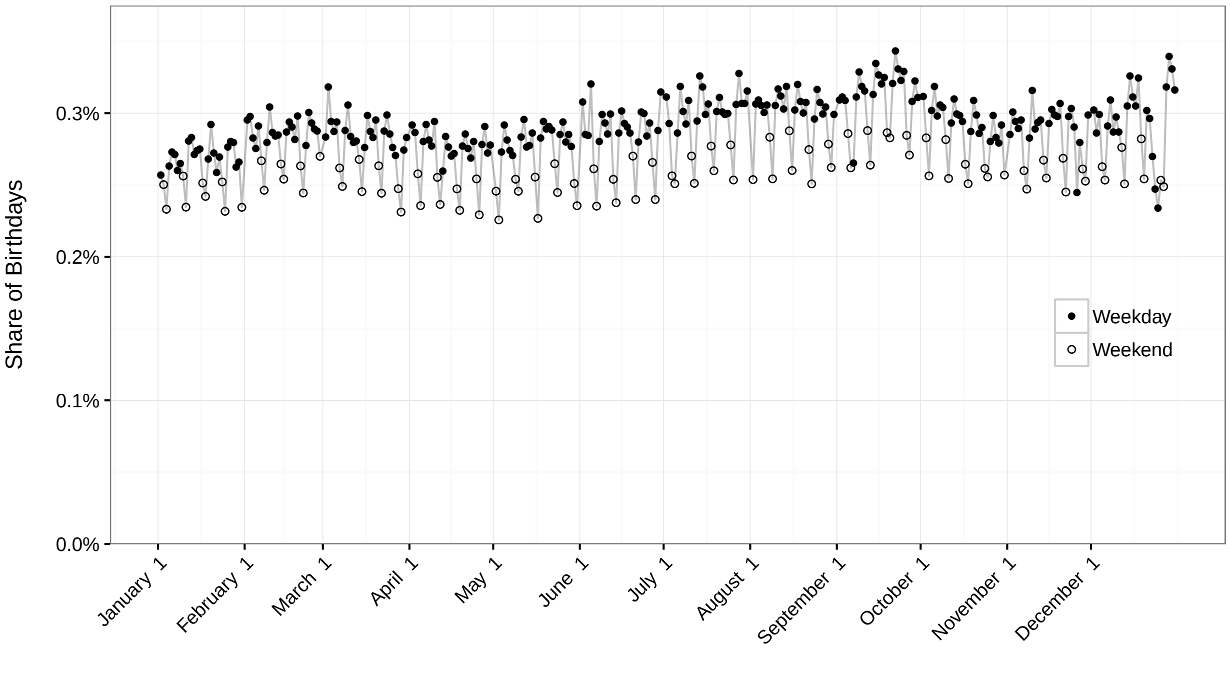
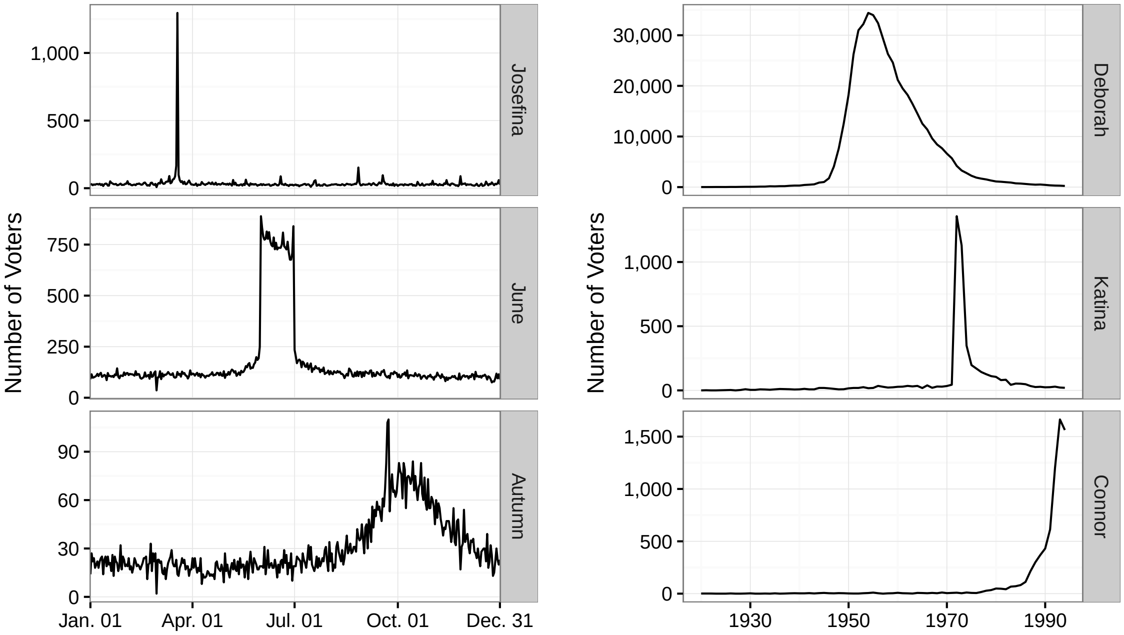


Figure 2: Distribution of (cleaned) birthdays in 1970 in the voter file.

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. And for those that are, it is logistically difficult to either request an absentee ballot or be physically present on Election Day in three or more 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.



(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.

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

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

where

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

and, for a smoothing parameter $\theta = 10,000$,

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

Our estimate of $\Pr(D = d)$ in Eq. (2) simply aggregates over all voters to generate the empirical birth day-of-week distribution. 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.

Section A.1 presents the formal 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 the well-known pattern 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.

This modeling strategy works well in most cases, but can lead to high-variance estimates for first names that are relatively rare, those that appear in our dataset fewer than 10,000 times. Consequently, in these cases we approximate $p_{b|f,l,y}$ by the empirical distribution of all birthdays in the year that the voter was born. In other words, for these rarer names, we set $\hat{p}_{b|f,l,y} = (\sum_{f'} C_{f',y,b})/(\sum_{b'} \sum_{f'} C_{f',y,b'})$.

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

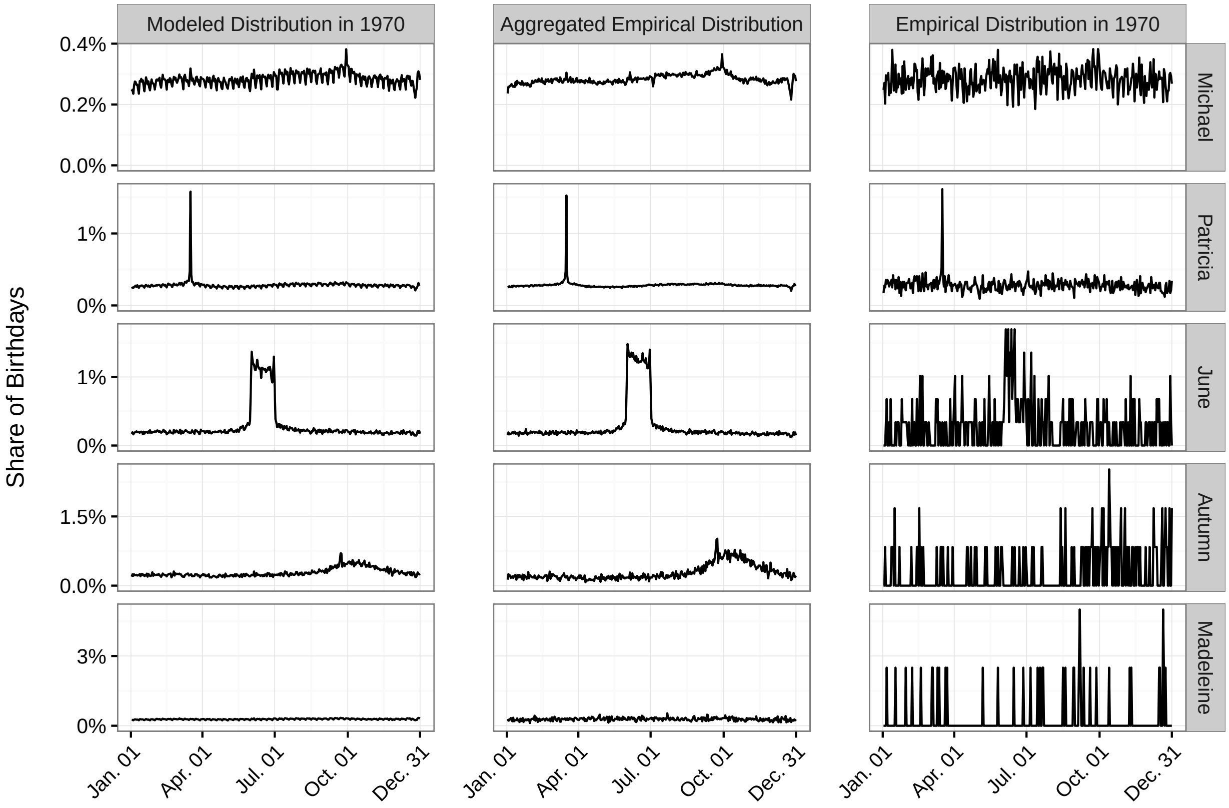


Figure 4: Modeled distribution of birthdays for voters born in 1970 for five different first names vs. the empirical distribution of birthdays for votes with those first names (aggregated across all years) and the empirical distribution of birthdays for voters with those first names born in 1970. Across all years (in 1970), we observe 1,669,961 (39,509) voters named Michael, 896,966 (7,616) voters named Patricia, 60,206 (295) voters named June, 10,813 (119) voters named Autumn, and 6,910 (40) 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, the empirical distribution of birthdays in the year is used.

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 5.1 *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

$$\hat{k}_{f,l,y} = \left(M_{f,l,y} - \binom{N_{f,l,y}}{2} \sum_i p^2_{b_i|f,l,y} \right) \Bigg/ \left(1 - \sum_i p^2_{b_i|f,l,y} \right). \tag{4}$$

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

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

The proof for Theorem 5.1 is included in the Appendix.

If we use all of the vote records in the voter file in our analysis, we find 2,892,574 pairs of vote records that match on first name, last name and date of birth among the 124,942,824 vote records observed in the 2012 election. If we assume that birthdays are uniformly distributed within the year, the theorem estimates the number of double votes as 1,953,393, while this estimate drops to 379,932 if we model the distribution using $\hat{p}_{b|f,l,y}$. But as discussed in the Data section, we have concerns about these estimates because they include observations with inaccurate dates of birth. 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 5.1 estimates the original number of double voters before the update happened, $k^{\mathrm{orig}}_{f,l,y}$, 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 5.1.

Lemma 5.1 *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,*

$$\hat{k}^{orig}_{f,l,y} = \frac{K_{f,l,y}}{(1-p_r)^2 - 2p_u} - \frac{N_{f,l,y}p_u}{(1+p_u-p_r+p_up_r)\left((1-p_r)^2-2p_u\right)} \tag{5}$$

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

The proof for Lemma 5.1 is included in the Appendix. Based on the lemma, if we have $p_r \gg p_u$ then the estimator simplifies to,

$$\hat{k}^{\text{orig}}_{f,l,y} \approx \frac{K_{f,l,y}}{1-2p_r} - \frac{N_{f,l,y}p_u}{(1-p_r)(1-2p_u)}. \tag{6}$$

Using Lemma 5.1 we can generalize the results of our analysis on the preferred sample to the whole population of 2012 voters in three steps.

1. *Exclude observations*: We drop vote records for registrations associated with a birthday on the first day of the month. This excludes 17,320,068 vote records from our analysis. We also drop vote records from the seven states (AK, DC, HI, MS, NH, WI, WY) with the highest rates of in-state duplicate votes for people living at the same address. This excludes an additional 3,416,475 votes, leaving our preferred sample of 104,206,281 vote records.

2. *Estimate rate of double voting for preferred sample*: Within our preferred sample of 104,206,281 vote records, there are 728,410 pairs of 2012 vote records that share the same first name, last name, and date of birth. Given our assumptions about $p_{b|f,l,y}$, Theorem 5.1 estimates that there were 19,977 (s.e. = 1,689) double votes in our preferred sample.[11]

[11] If we ignored the cyclic popularity of birthdays for different first names and instead assumed all voters

3. *Scaling to account for excluded observations*: While the FEC reported that 129,085,410 votes were cast in the 2012 presidential election, our preferred sample only includes 104,206,281 vote records. This implies that a vote record has about a 19.3% chance of being dropped. Our assumption is that each voter has the same probability of getting excluded from our analysis. If this assumption holds, Lemma 5.1 says we can generalize the rate of double voting in the broader population by multiplying our estimate by 1.54. This means that our estimate of the number of double votes in the whole population of voters is 30,765.

5.2 Accounting for Measurement Error

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. Such an 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 that a non-voter cast a ballot or that a voter abstained from voting. The remainder of this subsection extends the model from the previous section to incorporate the possibility that registrations not used to vote erroneously have a vote recorded in the voter file. The following subsection then presents evidence from an audit of Philadelphia poll books that is useful for assessing the rate at which this happens.

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

have the same distribution as the empirical distribution of birthdays of voters born in that year ($\hat{p}_{b|f,l,y} = \hat{\Pr}(B = b \mid Y = y)$), then our estimate would increase to 22,569 (s.e. = 1,684) double votes. Further assuming that birthdays were uniformly distributed within a year changes our estimate to 26,949 (s.e. = 1,678) double votes. If we didn't standardize first names, our estimate would decrease to 18,823 (s.e. = 1,571).

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, or 2.5% of, non-voting registration records 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 illusory double votes for every 100,000 voters that have a deadwood registration.[12]

We can use Lemma 5.1 to account for the potential measurement error in the recording of votes that we describe in the previous paragraph. 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 5.1, 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 $p_u = \frac{c(1-f_n)f_p}{N}$.[13] 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 5.1.

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 5.1

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

[13] 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 $\frac{c(1-f_n)f_p}{N}$.

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 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.[14] 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 $S_1, \ldots, S_u$. Since we do not expect duplicate registrations inside the state, we further assume that an observation and its copy cannot be in the same state. 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

[14] Pew (2012) reports that the vast majority of people with duplicate registrations are registered in only two states.

[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.

$\{(B_i, B_{i+m}) \mid 1 \leq i \leq k\}$ 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 5.2 provides an estimator for $c_{f,l,y}$ based on $T_{f,l,y}$.

Theorem 5.2 *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*

$$\hat{c}_{f,l,y} = \left(T_{f,l,y} - \left(\sum_{l=1}^{u} v_l \sum_{l=1}^{u} \bar{v}_l - \sum_{l=1}^{u} v_l \bar{v}_l \right) \sum_i p^2_{b_i|f,l,y} \right) \Bigg/ \left(1 - \sum_i p^2_{b_i|f,l,y} \right). \tag{7}$$

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

$$\mathrm{Var}(\hat{c}_{f,l,y}) \leq \left(\sum_{l=1}^{u} v_l \sum_{l=1}^{u} \bar{v}_l - \sum_{l=1}^{u} v_l \bar{v}_l \right)^2 \left[\frac{\sum_i p^2_{b_i|f,l,y}}{1 - \sum_i p^2_{b_i|f,l,y}} \right]. \tag{8}$$

The proof for Theorem 5.2 is included in the Appendix.

We observe $T = 1,792,170$ pairs in the voter file, and the theorem estimates that 1,563,603 (s.e. = 21,913) of these are duplicate registrations. If we assume $f_p = f_n = f$ as the clerical error rate, then we should plug in $p_u = \frac{1{,}563{,}603 \times f \times (1-f)}{104{,}206{,}281}$ and $p_r = f$ in Lemma 5.1 to correct for measurement error. After scaling the corrected estimate of double votes for records we dropped from our analysis using steps already 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.

Assessing Figure 5 requires knowledge of the clerical error rate in the population. We use the data collected from our Philadelphia poll book audit to better inform us about the rate at which registrations are incorrectly classified as having been used to vote in the voter file (i.e., a false positive). We find that 1.5% of registrations with an electronic voting record

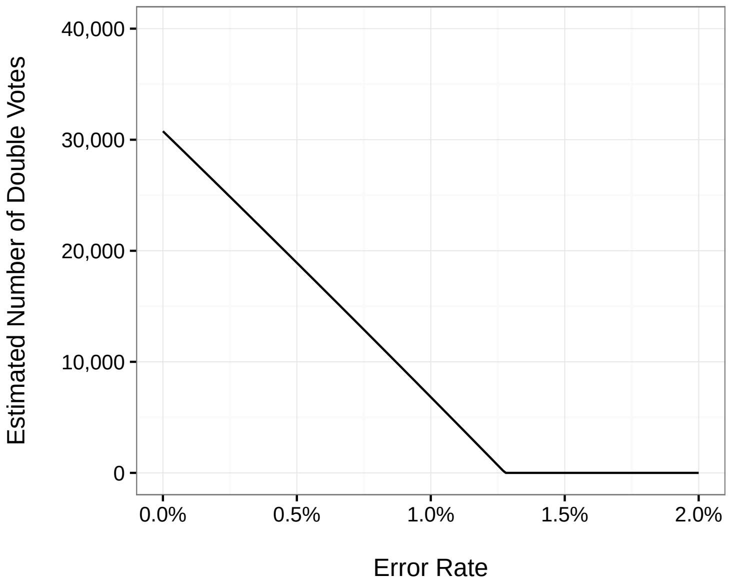


Figure 5: How the estimated number of double votes changes based on the clerical error rate.

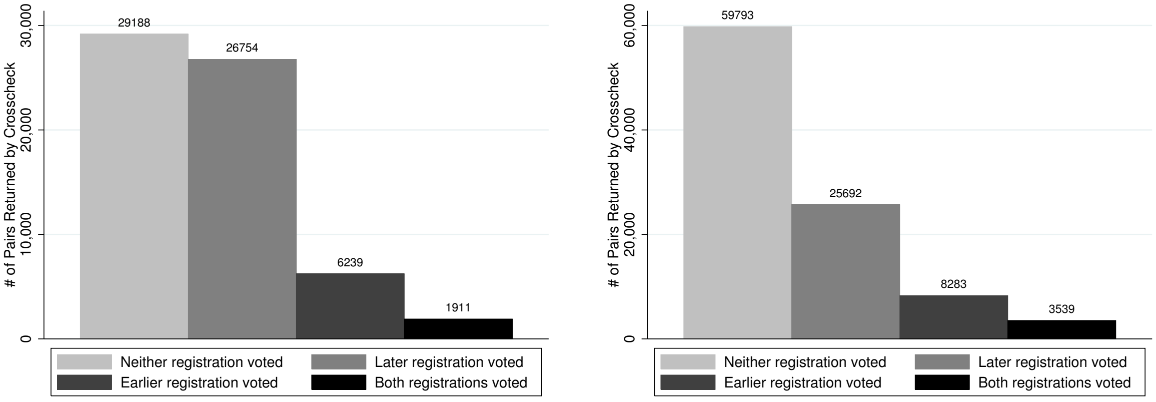
did not have a corresponding record of voting in a poll book. To investigate how many of these cases represent a false positive, we compare the past vote history of registrations with and without a discrepancy and conclude that measurement error likely explains a sizable portion, and possibly all, of the surplus double votes that we observe in the national voter file. Details on how we reach this conclusion, including a discussion of previous research, is presented in Section A.6 in the Appendix.

5.3 Crosscheck

Having found that the problem of double voting is, at worst, modest, we conclude by examining the tradeoffs involved in minimizing the potential for double votes to be cast. In particular, we examine how many registrations used to cast seemingly legitimate votes in Iowa might be impeded by attempts to purge the duplicate registrations necessary to cast two votes.

Figure 6 examines the turnout history of the potential double registrants that Crosscheck

Figure 6: Vote Records of Potential Double Registrations Identified by Crosscheck



2012

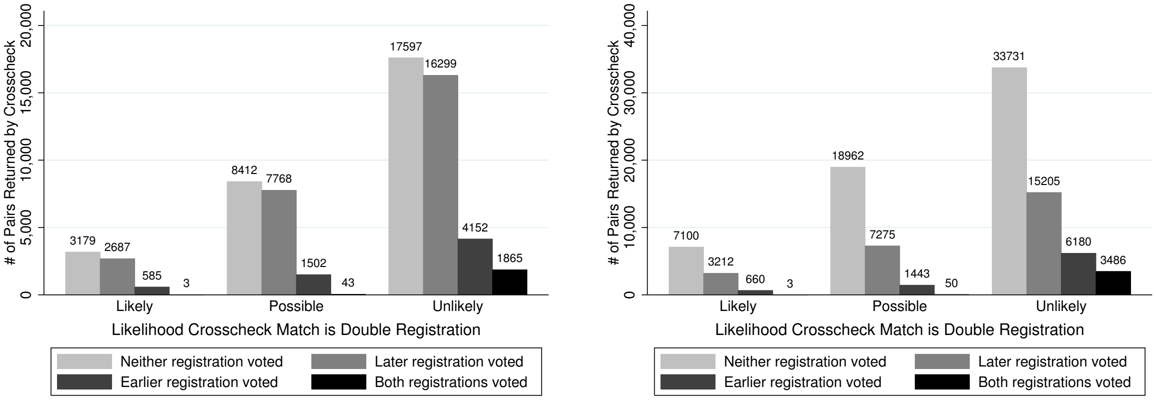
2014

identified in Iowa. The left panel shows that of the 64,092 pairs identified in mid-2012, there were 1,911, or about 3 percent, in which both registrations were used to vote later that same year. The right panel shows a similar trend two years later, in which about 3.6 percent of the 2014 pairs both had a record of voting.

While Crosscheck flagged all cases in which a registration in another state shared the same first name, last name, and date of birth as a registration in Iowa, the similarity of registrations' middle names and last four digits of social security numbers (SSN4s) is informative about the quality of any potential match. An important advantage of our Crosscheck records is that we are actually able to observe whether a matched pair shares a common SSN4. This additional information, which is not part of the TargetSmart voter file, allows us to better identify whether the potential double votes revealed in Figure 6 were actually cast by two distinct people with similar characteristics. A pairing in which both registration records have consistent middles names and common SSN4s is likely to be a case of the same person registered in two different states, while a pairing in which the registration records have inconsistent middles names or different SSN4s is likely to be a case in which there are two distinct people who live in different states with the same name and date of birth.

Figure 7 shows that most of the registration records Crosscheck identified as potential

Figure 7: Vote Records of Potential Double Registrations Identified by Crosscheck by Match Quality of Pairing



2012 2014

duplicates are likely for two distinct individuals in two different states with the same first name, last name, and date of birth. It disaggregates the data presented in Figure 6 based on the consistency of matched voters' middle names and SSN4s. Only 6,454, or about 10%, of the pairs identified in 2012 (left panel) had consistent middle names and common SSN4s. In about 30% of the pairings, no voter characteristics were inconsistent, but at least one of the records was missing a middle name and/or SSN4. The remaining 60% of the pairings returned by Crosscheck had inconsistent middle names and/or different SSN4s. The patterns were nearly identical in 2014 (right panel).

Among the likely double registrations, Figure 7 shows that there are almost no cases of double voting. There are only three cases in 2012 in which paired registration records had consistent middle names, common SSN4s, and both were used to cast a ballot. There are 43 more cases in which no voter characteristics were inconsistent, but at least one of the records was missing a middle name and/or SSN4.[16] The remaining 1,865 potential double votes in Figure 6 — comprising 98% of the potential double votes identified by Crosscheck in Iowa that year — came from pairings with inconsistent middle names and/or different

[16] Table A.2 and Table A.3 in the Appendix present the full distribution of cases by the quality of the middle name and SSN4 match in 2012 and 2014, respectively.

SSN4s. The results in 2014 are again nearly identical.

Not only would few double votes be eliminated by purging the voter registration rolls of the cases identified by Crosscheck, but many legitimate votes would be impeded by such a purge. Suppose a cautious election administrator engaging in voter list maintenance only looked at likely double registrations. There were 3,272 and 3,872 cases in 2012 and 2014, respectively, in which only one of the two registration records had a vote record attached to it. But it is not clear which of the two registrations is deadwood. For example, one might assume that the earlier registration would be obsolete. While the later registration is more likely to have a vote record than the earlier registration, there were 585 and 660 cases in 2012 and 2014, respectively, in which only the earlier registration had a vote record. Thus, even a purge of the voter registrations most likely to be duplicates risks eliminating registrations used to cast a substantial number of seemingly legitimate votes. We describe such votes as being impeded, because some of the affected people would recognize that their registration was at risk of being purged and take the necessary steps to fix the situation before Election Day. But others likely would not. And, at least in Iowa, 200 legitimate voters may be impeded from voting for every double vote stopped. This suggests the policies necessary to stop the relatively small number of double votes that do occur would put many more legitimate votes in jeopardy.

6 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 30,000, 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, if not all, 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, Ansolabehere, and Persily (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 names, last names, and dates of birth with registrations used to vote in other states. Given our analysis of Crosscheck data from Iowa, it is unlikely that more than a handful of the 35,750 actually are the same people and double voted.

Media stories, like that referenced above, suggest that Crosscheck is being irresponsible in the way that it disseminates information back to participating states. Crosscheck returns all cases where a registration in one state shares the same first name, last name, and date of birth as a registration in another state, even when other information suggests that the registrations are from two distinct individuals. As a result, Crosscheck was able to circulate Figure A.4 in 2013, which reported that it had identified 1,395,074 "potential duplicate voters" among the 15 states participating in the previous year, including 100,140 in Iowa. But our results suggest that fewer than 10 of these potential duplicate voters in Iowa actually double voted. Most Crosscheck matches are not double registrations, and few of the likely double registrations appeared to double vote. While Crosscheck's 2014 participation guide does acknowledge that, with their system, "a significant number of apparent double votes are false positives," it still distributes information in such a way that encourages misleading news stories. Given that a belief in free and fair elections is critical for democracy, it is problematic when programs like Crosscheck distribute information in such a way that it is likely to cause people to overstate the potential problem of voter fraud.

Crosscheck may justify the provision of these data on the grounds that providing more information is always better than providing less. But not all election administrators are able to discern the quality of a given Crosscheck match. For example, in a 2013 letter to local election officials, the Virginia State Board of Elections mischaracterized the state's Crosscheck results as "extremely accurate" and wrongly described any match as "based on a 100% exact match of [voters'] first name, last name, date of birth and last four digits of their Social Security Number."[17] The same letter offered an interpretation allowing local registrars to sidestep the typical federal process outlined in the Help America Vote Act where administrators mail confirmation notices prior to purging any registration and instead immediately cancel them based solely on the Crosscheck results.[18] The decentralized nature of election administration adds yet another opportunity for Crosscheck results to be misused. In 2014, for example, election officials in Ada County, Idaho mistakenly revoked all of the registrations flagged by Crosscheck, unbeknownst to the state, or even the affected voters. Among the purged registrations was a school district superintendent, who learned of the mistake when she attempted to vote for a school bond she supported (Sewell, 2014). Our results show that such a careless use of Crosscheck data across the country could disenfranchise enough people to affect electoral outcomes.

More broadly, this paper suggests that both Democrats and Republicans need to acknowl-

[17] This letter was entered into evidence as part of a lawsuit brought by the state Democratic party. See *Democratic Party of Virginia v. Virginia State Board of Elections*, 2013 WL 5741486.

[18] Using an FAQ format, the letter asked whether "being listed in the [Crosscheck] report [is] a sufficient basis for cancellation," to which the response was "Legally yes." The letter interpreted all the matches returned by Crosscheck as an official notice from another jurisdiction and cited Va. Code §24.2-427(B)(iv) which notes that "The general registrar shall cancel the registration of... (iv) all persons for whom a notice has been received... from the registration official of another jurisdiction that the voter has registered to vote outside the Commonwealth." This interpretation was also confirmed in a phone call with non-political appointees in the Virginia Department of Elections.

edge that there are tradeoffs between accessibility and integrity in electoral administrative policy. In the case of Crosscheck, even a relatively conservative use of their data would impede 200 legal votes, based on the data from Iowa, 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.*

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.

Bartels, Larry M. (2002). “Beyond the running tally: Partisan bias in political perceptions.” *Political Behavior* 24(2): 117–150.

Baum, Dale (1991). “Pinpointing apparent fraud in the 1861 Texas secession referendum.” *Journal of Interdisciplinary History* 22(2): 201–221.

Beber, Bernd and Scacco, Alexandra (2012). “What the numbers say: A digit-based test for election fraud.” *Political Analysis* 20(2): 211–234.

Bentele, Keith G and O’Brien, Erin E (2013). “Jim Crow 2.0? Why states consider and adopt restrictive voter access policies.” *Perspectives on Politics* 11(4): 1088–1116.

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.

Caro, Robert A. (1990). *Means of ascent: The years of Lyndon Johnson II.* New York, NY: Vintage Books.

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.

Fukumoto, Kentaro and Horiuchi, Yusaku (2011). “Making outsiders’ votes count: Detecting electoral fraud through a natural experiment.” *American Political Science Review* 105(3): 586–603.

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.

Hicks, William D. et al. (2015). “A principle or a strategy? Voter identification laws and partisan competition in the American states.” *Political Research Quarterly* 68(1): 18–33.

Hohmann, James (2014). “Walker defends voter ID law in debate.” *Politico* October 11. URL: http://www.politico.com/story/2014/10/scott-walker-wisconsin-voter-id-law-111804.

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. URL: http://www.nationalreview.com/corner/374882/nc-state-board-finds-more-35k-incidents-double-voting-2012-andrew-johnson.

King, Ronald F (2001). “Counting the votes: South Carolina’s stolen election of 1876.” *Journal of Interdisciplinary History* 32(2): 169–191.

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. URL: http://www.kansas.com/news/politics-government/article39670275.html.

Mayer, Jane (2012). "The voter-fraud myth." *The New Yorker* October 29. URL: http://www.newyorker.com/magazine/2012/10/29/the-voter-fraud-myth.

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. URL: http://thehill.com/opinion/dick-morris/203019-dick-morris-investigate-2012-vote-fraud.

Noble, Jason (2013). "Iowa voter fraud probe nets few cases, No trials since July 2012." *Des Moines Register* December 16. URL: http://www.desmoinesregister.com/story/news/politics/2013/12/16/voterfraud/4037057/.

Pew (2012). *Inaccurate, costly, and inefficient: Evidence that America's voter registration system needs an upgrade.* Pew Charitable Trusts. URL: http://www.pewtrusts.org/en/

research-and-analysis/reports/2012/02/14/inaccurate-costly-and-inefficient-evidence-that-americas-voter-registration-system-needs-an-upgrade.

Powell, Lawrence N (1989). “Correcting for fraud: a quantitative reassessment of the Mississippi ratification election of 1868.” *Journal of Southern History* 55(4): 633–658.

Sewell, Cynthia (2014). “Ada County mistakenly revokes 765 voter registrations - Staff failed to take steps to confirm duplicate sign-ups before making cancellations.” *Idaho Statesman* August 29.

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. URL: https://www.washingtonpost.com/politics/for-trump-a-new-rigged-system-the-election-itself/2016/08/02/d9fb33b0-58c4-11e6-9aee-8075993d73a2_story.html.

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:*

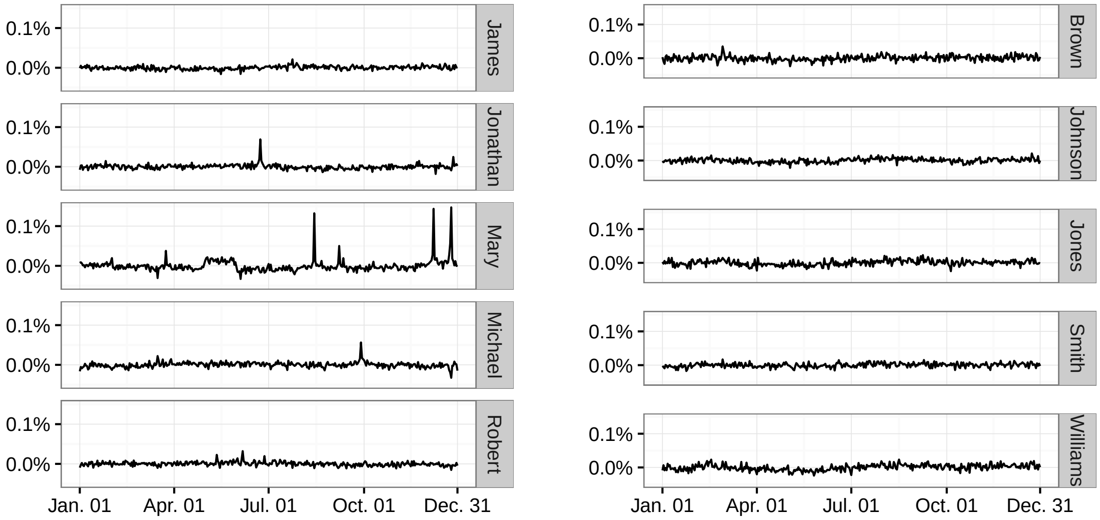
1. *If* $d_{b,y_1} = d_{b,y_2} \forall b$, *then* $\Pr(B = b \mid Y = y_1, F = f) = \Pr(B = b \mid Y = y_2, F = f)$*;*

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,

$$\Pr(B = b \mid F = f, Y = y) = \frac{Pr(B = b \mid F = f)\ \Pr(D = d_{b,y})}{\sum_{b'} \Pr(B = b' \mid F = f)\ \Pr(D = d_{b',y})}\ . \tag{9}$$



(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,

$$\Pr(B = b \mid Y = y, F = f) = \frac{\Pr(B = b, D = d_{b,y} \mid F = f)}{\sum_{(y' \text{ s.t. } y' \sim y)} \Pr(Y = y' \mid F = f)}. \tag{10}$$

Using Bayes' rule, we can rewrite the numerator in Eq. (10) as,

$$\begin{aligned} \Pr(B = b, D = d_{b,y} \mid F = f) &= \frac{\Pr(F = f, D = d_{b,y} \mid B = b) \ \Pr(B = b)}{\Pr(F = f)} \\ &= \frac{\Pr(F = f \mid B = b) \ \Pr(D = d_{b,y} \mid B = b) \ \Pr(B = b)}{\Pr(F = f)} \end{aligned} \tag{11}$$

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,

$$\Pr(F = f \mid B = b) = \frac{\Pr(B = b \mid F = f) \ \Pr(F = f)}{\Pr(B = b)}. \tag{12}$$

Plugging Eq. (12) into Eq. (11) and simplifying gives us that

$$\begin{aligned} \Pr(B = b, D = d_{b,y} \mid F = f) &= \Pr(F = f \mid B = b) \times \Pr(D = d_{b,y} \mid B = b) \times \frac{\Pr(B = b)}{\Pr(F = f)} \\ &= \frac{\Pr(B = b \mid F = f) \ \Pr(F = f)}{\Pr(B = b)} \times \Pr(D = d_{b,y} \mid B = b) \times \frac{\Pr(B = b)}{\Pr(F = f)} \\ &= \Pr(B = b \mid F = f) \ \Pr(D = d_{b,y} \mid B = b) \\ &= \Pr(B = b \mid F = f) \ \Pr(D = d_{b,y}) \end{aligned} \tag{13}$$

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. (13) into the numerator of Eq. (10) gives us that

$$\begin{aligned}
\Pr(B = b \mid Y = y, F = f) &= \frac{\Pr(B = b, D = d_{b,y} \mid F = f)}{\sum_{(y' \text{ s.t. } y' \sim y)} \Pr(Y = y' \mid F = f)} \\
&= \frac{\Pr(B = b \mid F = f)\ \Pr(D = d_{b,y})}{\sum_{(y' \text{ s.t. } y' \sim y)} \Pr(Y = y' \mid F = f)} \\
&= \frac{\Pr(B = b \mid F = f)\ \Pr(D = d_{b,y})}{Z(f, y)}. \qquad (14)
\end{aligned}$$

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,

$$Z(f, y) = \sum_{b'} \Pr(B = b' \mid F = f)\ \Pr(D = d_{b',y}). \qquad (15)$$

Plugging in Eq. (15) to Eq. (14) yields the lemma.

A.2 Proof of Theorem 5.1

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 $1 \le i < j \le m + k$, let $A_{i,j}$ indicate whether $B_i = B_j$. Then by the linearity of expectation,

$$\mathbb{E}M = \mathbb{E}\left(\sum_{1 \le i < j \le m+k} A_{i,j}\right) = \sum_{1 \le i < j \le m+k} \mathbb{E}A_{i,j}. \qquad (16)$$

For $1 \le i \le k$, $\mathbb{E}A_{i,m+i} = 1$ 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

$$\mathrm{Var}(M) = \sum_{1 \le i < j \le m+k} \mathrm{Var}(A_{i,j}) + 2 \sum_{S} \mathrm{Cov}(A_{i,j}, A_{k,l}) \tag{17}$$

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,

$$\mathrm{Var}(A_{i,j}) = \mathbb{E}A_{i,j} - (\mathbb{E}A_{i,j})^2 . \tag{18}$$

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,

$$\sum_{1 \le i < j \le m+k} \mathrm{Var}(A_{i,j}) = \left(\binom{m+k}{2} - k \right) \left(\sum_s p_s^2 - \left(\sum_s p_s^2 \right)^2 \right) . \tag{19}$$

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,

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

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,

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

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

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

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

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

A.3 Proof of Lemma 5.1

To simplify the notation, we represent $K_{f,l,y}$ by K, $N_{f,l,y}$ by N, and $k_{f,l,y}^{\mathrm{orig}}$ 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,

$$\mathbb{E}K = k^{\text{orig}}(1-p_r)^2 + (n - 2k^{\text{orig}})p_u = k^{\text{orig}}\left[(1-p_r)^2 - 2p_u\right] + np_u. \tag{20}$$

Rearranging terms, we get,

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

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 $(1-p_u)(1-p_r) = 1 - p_u - p_r + p_u p_r$. Therefore,

$$\mathbb{E}N = \sum_{i=1}^{n} 2p_u + 1 - p_u - p_r + p_u p_r = n(1 + p_u - p_r + p_u p_r) \tag{22}$$

Substituting $n = \frac{\mathbb{E}N}{1+p_u-p_r+p_up_r}$ into the Eq. (21), we have $\mathbb{E}\hat{k}^{\text{orig}} = k^{\text{orig}}$.

Note that in the proof of Theorem 5.1 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 5.2

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,

$$\mathbb{E}T = \mathbb{E}\left(\sum_{(B_i,B_j)\in\mathcal{Q}} A_{i,j}\right) = \sum_{(B_i,B_j)\in\mathcal{Q}} \mathbb{E}A_{i,j}. \tag{23}$$

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

$$\mathrm{Var}(T) = \sum_{(B_i,B_j)\in\mathcal{Q}} \mathrm{Var}(A_{i,j}) + 2\sum_{S} \mathrm{Cov}(A_{i,j}, A_{k,l}) \tag{24}$$

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,

$$\mathrm{Var}(A_{i,j}) = \mathbb{E}A_{i,j} - \left(\mathbb{E}A_{i,j}\right)^2 . \tag{25}$$

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

$$\sum_{(B_i,B_j)\in\mathcal{Q}} \mathrm{Var}(A_{i,j}) = (|\mathcal{Q}| - c)\left(\sum_s p_s^2 - \left(\sum_s p_s^2\right)^2\right). \tag{26}$$

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

$$\mathrm{Cov}(A_{i,j}, A_{k,l}) \le \sqrt{\mathrm{Var}(A_{i,j})\mathrm{Var}(A_{k,l})}. \tag{27}$$

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

$$\sum_S \mathrm{Cov}(A_{i,j}, A_{k,l}) \le \binom{|\mathcal{Q}| - c}{2}\left(\sum_s p_s^2 - \left(\sum_s p_s^2\right)^2\right). \tag{28}$$

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

$$\mathrm{Var}(T) \le (|\mathcal{Q}| - c)^2\left(\sum_s p_s^2 - \left(\sum_s p_s^2\right)^2\right). \tag{29}$$

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 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, last name, suffix, date of birth, and turnout history associated with a voter registration.[19] These data provide a nearly comprehensive list of 2012 general election participation: the data include 124,942,824 vote records from the 2012 election, as compared to the 129,085,410 votes cast for a presidential candidate nationwide.[20] 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 be always be reported accu-

[19] 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.

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

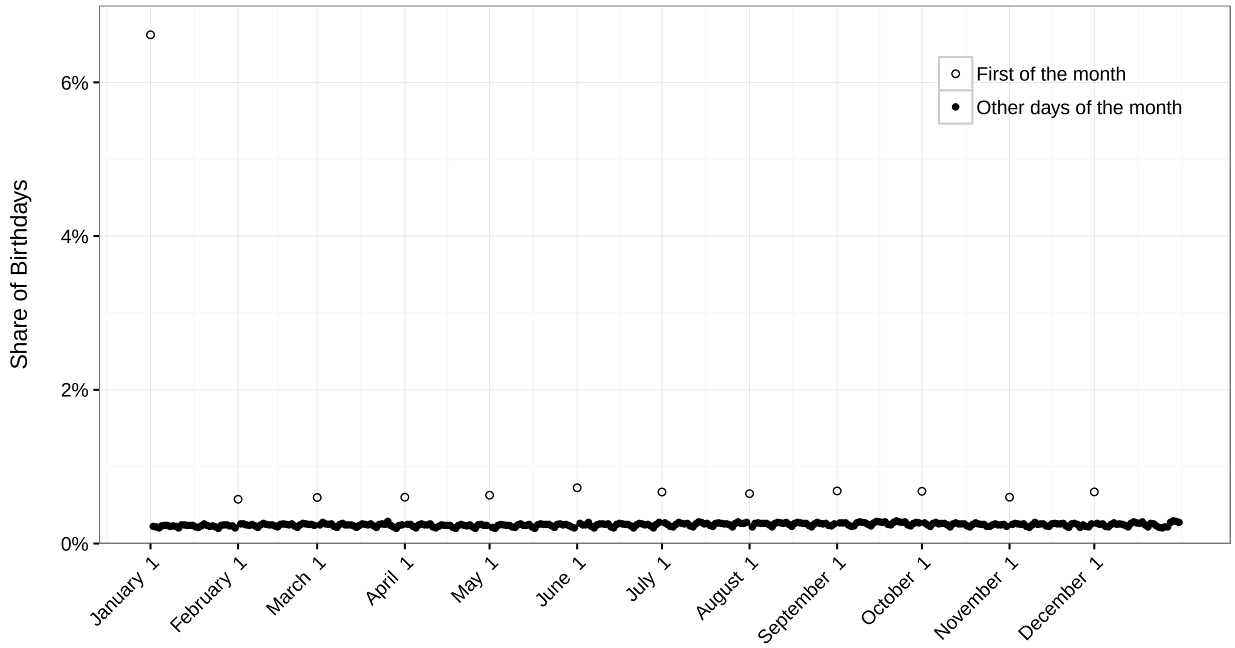


Figure A.2: Distribution of birthdays in 1970 in the voter file.

rately in the voter file. Figure A.2 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.[21] Such 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 7,430 and 2,318 in-state duplicate voters in Mississippi and Wisconsin, respectively. In a vast majority

[21] 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 show a higher concentration of certain voter registration birthdays

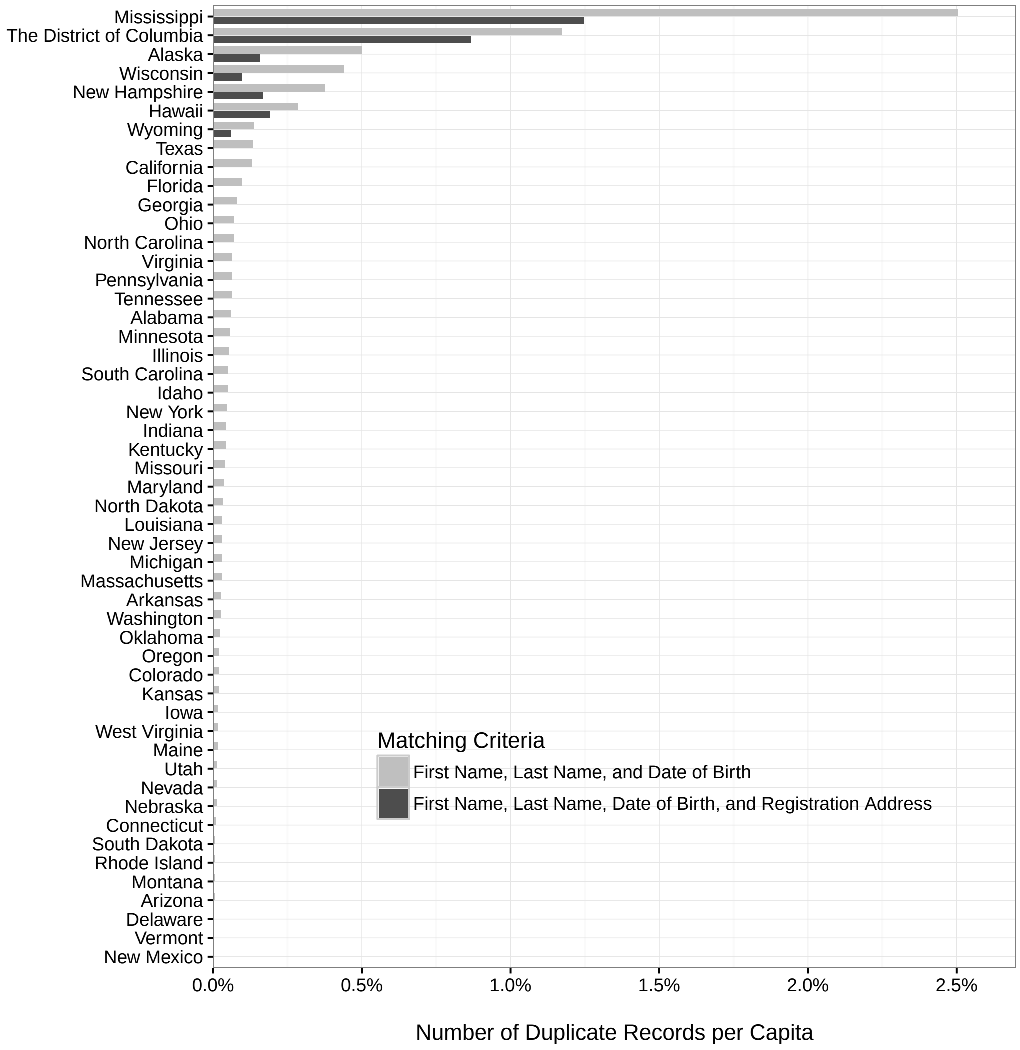


Figure A.3: Distribution of potential multi-generational matches within a state.

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.3 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 seemingly double voters whose multiple registrations list the same household.

A.6 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 better inform us about the rate at which registrations are incorrectly classified as having been used to vote in the voter file. We 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). We found another 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). Given that we audited 11,663 vote records, our estimated rate at which registrations are wrongly assigned electronic voting records is 1.5% if all of these cases represent a false positive.

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 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 67% of 2010 voters also turned out in 2006, while only 19% 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 45% 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 1.5%.

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.[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

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

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	0.422***	0.369***	0.427***	0.232***
	(0.006)	(0.005)	(0.006)	(0.004)
Signature error	−0.161***	−0.138***	−0.073**	−0.129***
	(0.035)	(0.031)	(0.035)	(0.025)
Registration error	−0.348***	−0.363***	−0.089	−0.179***
	(0.077)	(0.069)	(0.078)	(0.055)
Constant	0.185***	0.118***	0.468***	0.042***
	(0.019)	(0.017)	(0.019)	(0.014)
R^2	0.244	0.230	0.253	0.153

Note: N = 25,641 registered voters in the 47 precincts that were audited.
Unreported precinct fixed effects included in all models.
*p<0.1; **p<0.05; ***p<0.01.

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 all, of the surplus double votes that we observe in the national voter file.

A.7 Additional Figures

A.8 Additional Tables

Table A.2: Crosscheck Matches in Iowa in 2012 by Quality of Match

		2012 Turnout of Matched Pair			
			Later	Earlier	
Quality of Match	N	Neither	Reg. Date	Reg. Date	Both
SSN4s Match and Middle Names are Consistent (Likely)	6454	3179	2687	585	3
SSN4s Match and At Least 1 Middle Name is Unknown to Crosscheck (Possible)	314	194	94	26	0
SSN4s Match and Middle Names are Inconsistent (Unlikely)	9726	4803	3899	1022	2
At Least 1 SSN4 is Unknown to Crosscheck and Middle Names are Consistent (Possible)	16970	7979	7524	1434	33
At Least 1 SSN4 and Middle Name is Unknown to Crosscheck (Possible)	441	239	150	42	10
SSN4s Match and Middle Names are Inconsistent (Unlikely)	26206	11397	11172	2503	1134
SSN4s Don't Match and Middle Names are Consistent (Unlikely)	915	427	373	97	18
SSN4s Don't Match and At Least 1 Middle Name is Unknown to Crosscheck (Unlikely)	55	35	11	4	5
SSN4s Don't Match and Middle Names are Inconsistent (Unlikely)	3011	935	844	526	706

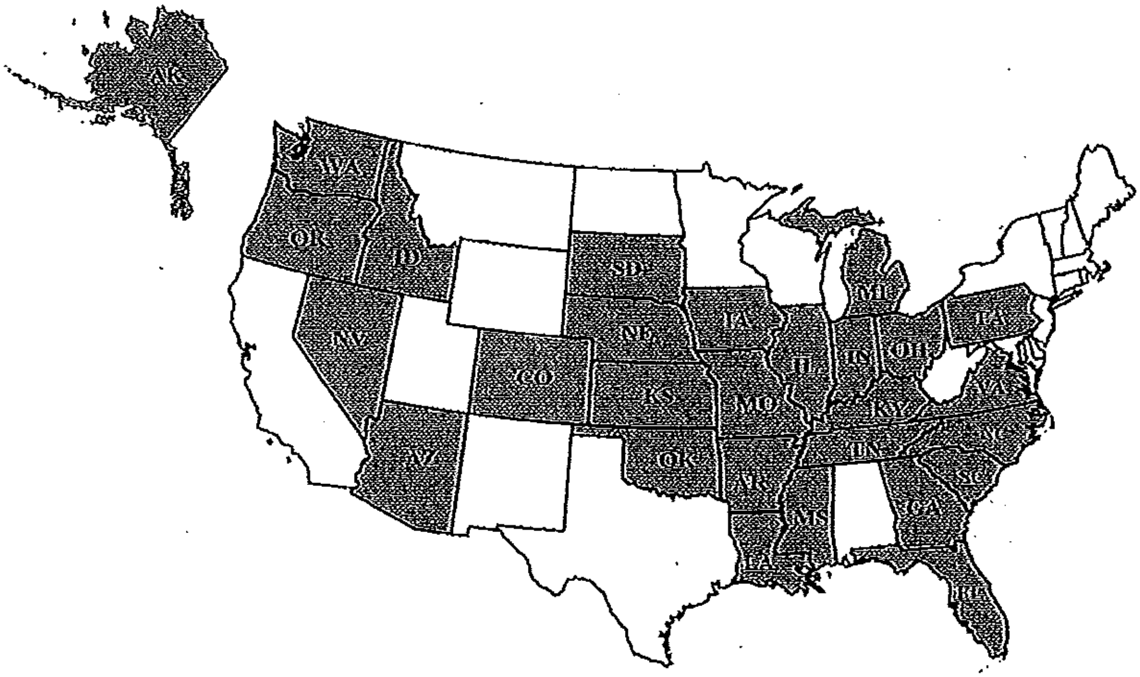
A.9 Crosscheck 2014 Participation Guide

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, <u>and</u>
   b. One or both of the following data elements match:
      - last four digits of Social Security number <u>and</u>
      - 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

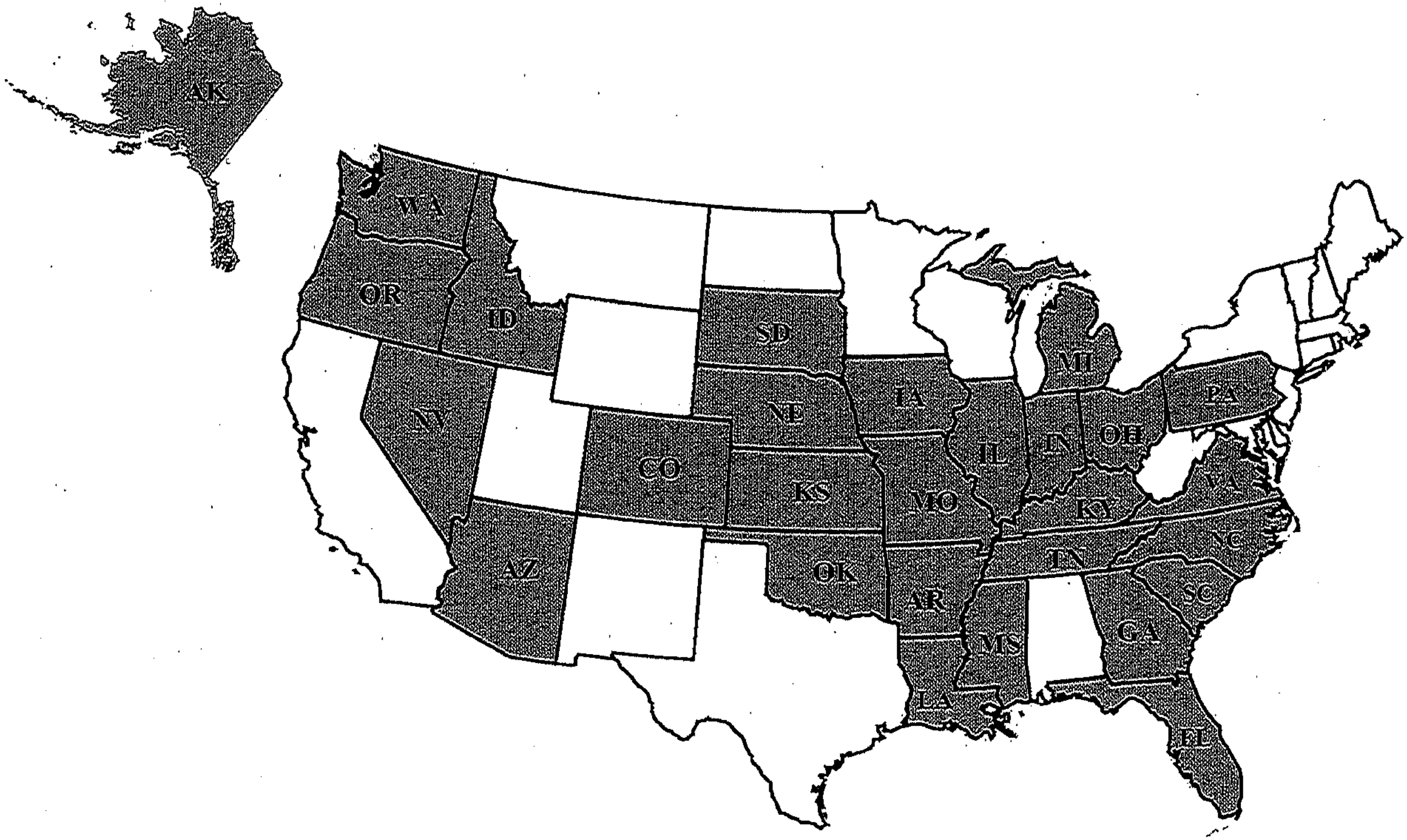
1) Download your state's self decrypting .exe file from the FTP site
2) Double click the file.
3) Enter the passphrase
   i) This will decrypt the file
   ii) The resulting .zip file contains 2 folders and a spreadsheet
4) Extract all files
5) Please delete your state's file from the FTP site once you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

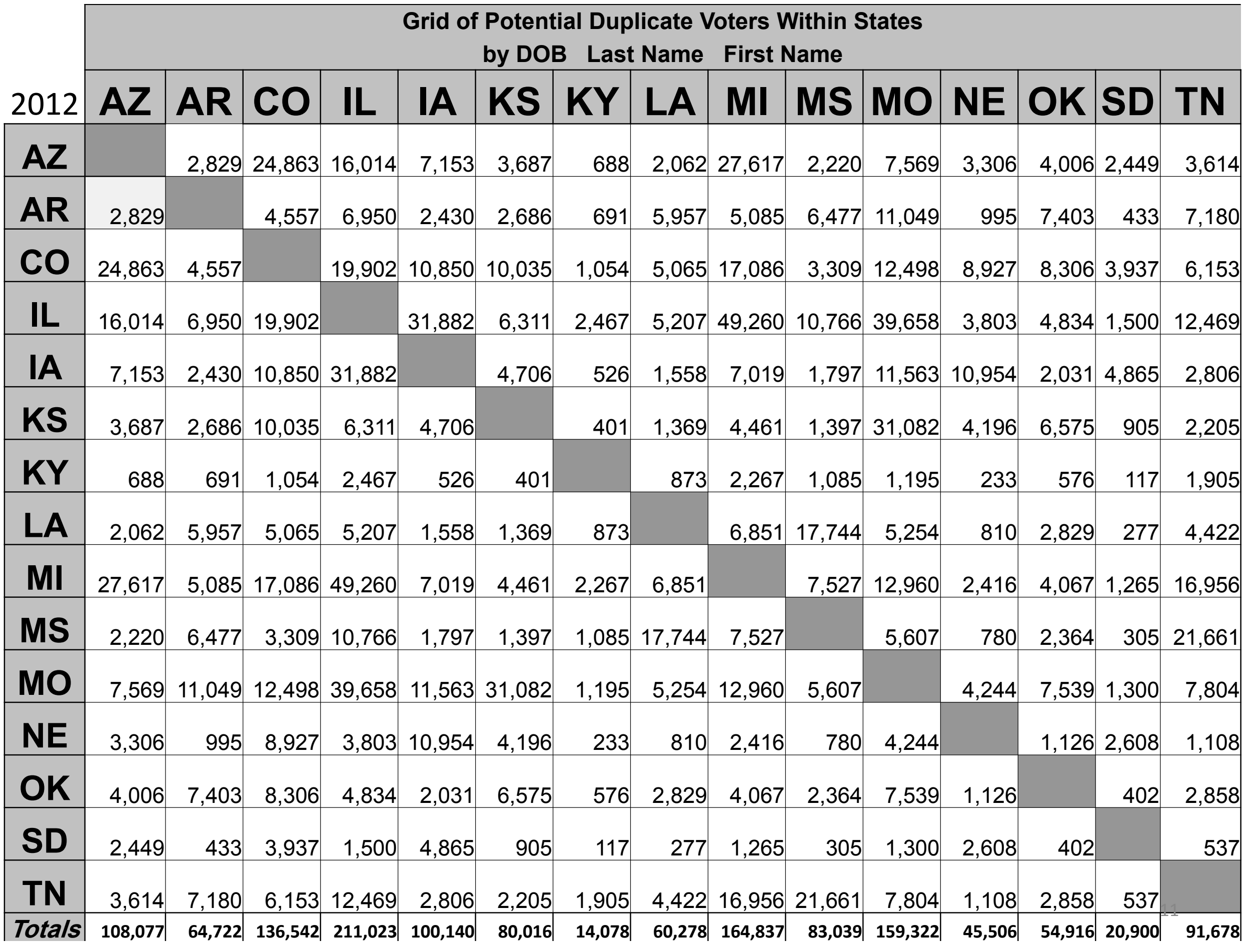
- ❖ One folder, SingleRowOutput:
  - ➢ Contains one result file with the potential match count of each comparison and total for that BaseState
  - ➢ Contains individual result files for your state compared with each other participating state
    - ▪ Ideally opened programmatically
  - ➢ Contains one 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



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

Table A.3: Crosscheck Matches in Iowa in 2014 by Quality of Match

		2012 Turnout of Matched Pair			
			Later	Earlier	
Quality of Match	N	Neither	Reg. Date	Reg. Date	Both
SSN4s Match and Middle Names are Consistent (Likely)	10975	7100	3212	660	3
SSN4s Match and At Least 1 Middle Name is Unknown to Crosscheck (Possible)	622	494	108	20	0
SSN4s Match and Middle Names are Inconsistent (Unlikely)	11971	7911	3309	747	4
At Least 1 SSN4 is Unknown to Crosscheck and Middle Names are Consistent (Possible)	26323	17875	7031	1382	35
At Least 1 SSN4 and Middle Name is Unknown to Crosscheck (Possible)	785	593	136	41	15
SSN4s Match and Middle Names are Inconsistent (Unlikely)	36116	21741	9813	3116	1446
SSN4s Don't Match and Middle Names are Consistent (Unlikely)	1842	1162	460	182	38
SSN4s Don't Match and At Least 1 Middle Name is Unknown to Crosscheck (Unlikely)	171	83	32	33	23
SSN4s Don't Match and Middle Names are Inconsistent (Unlikely)	8502	2834	1591	2102	1975