CALTECH/MIT
VOTING TECHNOLOGY PROJECT
A multi-disciplinary, collaborative project of
the California Institute of Technology – Pasadena, California 91125 and
the Massachusetts Institute of Technology – Cambridge, Massachusetts 02139

EVALUATING THE QUALITY OF CHANGES IN VOTER REGISTRATION DATABASES

Seo-young Silvia Kim, California Institute of Technology
Spencer Schneider, California Institute of Technology
R. Michael Alvarez, California Institute of Technology

Key words:*voter registration data*

***American Politics Research,* Forthcoming**

VTP WORKING PAPER #134

Evaluating the Quality of Changes in Voter Registration Databases

Seo-young Silvia Kim[*†]   Spencer Schneider
R. Michael Alvarez
California Institute of Technology

August 1, 2019

Abstract

The administration of elections depends crucially upon the quality and integrity of voter registration databases. In addition, political scientists are increasingly using these databases in their research. However, these databases are dynamic, and may be subject to external manipulation and unintentional errors. In this paper, using data from Orange County, California, we develop two methods for evaluating the quality of voter registration data as it changes over time: (1) generating audit data by repeated record linkage across periodic snapshots of a given database, and monitoring it for sudden anomalous changes; and (2) identifying duplicates via an efficient, automated duplicate detection, and tracking new duplicates and deduplication efforts over time. We show that the generated data can serve not only to evaluate voter file quality and election integrity, but also as a novel source of data on election administration practices.

[*] We thank the John Randolph and Dora Haynes Foundation for providing funding to support the research reported in this paper. Schneider thanks the Caltech Summer Undergraduate Research Fellowship program for supporting his work on this project. We thank Neal Kelley and his team at the Orange County Registrar of Voters (OCROV) office for their partnership on this project. We especially thank Justin Berardino of the OCROV for answering our many questions, and facilitating the transfer of the data we use in this paper. We also thank the two anonymous referees and the participants of ESRA 2019 and Polmeth 2019, especially Jamie Monogan, for their constructive feedback.

[†] The R package containing the algorithm codes are in GitHub repository `sysilviakim/voterdiffR`. Implementation for the Orange County data is under GitHub repository `monitoringtheelection/auditingVR`. For questions on the code, please file an issue.

Voter files are an important resource for political research, and also crucial for the integrity of the administration of elections, as they dictate who votes. The files constantly change—but not all changes are intentional or welcome. External intrusions into voter files became salient issues in the 2016 presidential election (Sanger, 2018). Media and federal officials reported that foreign actors attempted to access voter data in various states, and that they may have tried to alter registration data to manipulate election outcomes or undermine public trust.[1] On the other hand, internal quality deterioration can sometimes occur because voter files are large, dynamic, and complex. For example, in the 2018 June primary in California, records for many as 77,000 in the state's system were duplicated inadvertently by the Department of Motor Vehicles; in the 2018 primary election in Los Angeles County, 118,000 voters were left off precinct rosters due to a merge error.[2]

While election officials work tirelessly to guard against cyberattack and human error, there are calls for independent auditing of voter files, and generally to improve their quality (Alvarez et al., 2005, 2009; Ansolabehere & Hersh, 2010). Past studies of voter data quality have been static in their focus, as in Ansolabehere & Hersh (2014).[3] However, as we have seen, data quality can sharply change over time, presenting problems both to election administrators and to scholars using the data. In this paper, we present two methods that evaluate the internal validity of voter registration data as it changes over time, which increases assurance of voter file quality and provides interesting data for election scholars, as a novel source of data on election administration practices.

Collaborating with the Orange County Registrar of Voters (OCROV), our first approach matches voter *snapshots* at different points in time, and quantifies the changes to the file. As voter files are dynamic, some rate of change is expected, due to new registrations, residential mobility, deceased

[1] Perlroth, N., Wines, M., & Rosenberg, M. (Sep 1, 2017). Russian Election Hacking Efforts, Wider Than Previously Known, Draw Little Scrutiny. *The New York Times*; Fandos, N., & Wines, M. (May 8, 2018). Russia Tried to Undermine Confidence in Voting Systems, Senators Say. *The New York Times*.

[2] Myers, J. (2018, May 24). One voter, two registration forms: Errors reported in rollout of California's motor voter system. *Los Angeles Times*; Reyes, E. A., & Smith, D. (2018, June 6). Officials demand answers after more than 118,000 people were left off L.A. County voter roster. *Los Angeles Times*.

[3] We will show that besides the static vs. dynamic evaluation, there are two more differences between previous work like Ansolabehere & Hersh (2014) and ours. One is that our data is contemporary, and we are developing methods to help election administrators understand the quality of data that in California is dynamic, and changable, as different governmental agencies have access to and can manipulate registration data. The second is that we use an exact copy of the data that OCROV uses to administer elections in Orange County; Ansolabehere & Hersh (2014) used data from a third-party vendor, and it is likely that such data was subjected to data cleaning and manipulation by the vendor.

voters, and changes in personal information. The resulting time-series of changes can undergo statistical anomaly detection to find anomalous changes, which represent those that depart from the expected rate of change. While this approach can provide notification of a sudden deterioration of database quality, we argue that the generated audit data can also be of scholarly interest, serving as important source of information on election administration, a rare window into this important—and often overlooked—component of the democratic process. Our second approach is a duplication detection scheme which provides a list of potential duplicates with a principled, automated approach, while minimizing cases where the election official might accidentally delete a valid, non-duplicate voter. Combined with voter file information on how the registration data was generated, we show that we can determine the origin of duplicate records, and track incoming duplicates and deduplication efforts over time, which is another key metric of dynamic data quality.

Methods and Data

To quantify the data changes and to detect duplicates, we first need to decide which records correspond to the same person. For the former, we need to decide whether a voter A in yesterday's data and a voter A' in today's data are the same person, and for the latter, we need to decide whether voters A' and A'' that simultaneously exist are the same people. This is called *entity resolution* or *record linkage*. We take a unique approach by using record linkage on 252 daily "snapshots" of the voter file from Orange County, from April 26, 2018 to May 24, 2019. Because exact matching often fails due to typos, we use probabilistic record linkage, largely established by Fellegi & Sunter (1969). It assumes a latent variable of true match status for the two records being compared. The latent status will generate different distributions of agreement levels in each field such as names. Less frequent values provide more distinguishing power, since the odds ratio of two records with rare values being a match against being a nonmatch is higher than with frequent values—for instance, two records with surname 'Quoss' are more likely a true match than with 'Smith.'

Often with a computationally simplifying assumption of conditional independence, the parameters, which are field agreements conditional on match status, are estimated via maximum likelihood by the Expectation-Maximization (EM) algorithm. If a composite odds ratio that combines all the odds ratio from each field falls below a lower threshold, the records are deemed non-matches, if above an upper threshold, deemed matches—if in between, a clerical review is required. For a graphical representation of this idea, see Supplementary Materials Section 3.

However, because for every record pair the agreement level between each field has to be calculated—especially a continuous string distance if the field is non-numeric—this is computationally intensive. For cost reduction, *blocking* can be employed, i.e., only performing comparisons based on value agreement of a certain field; if we block by the date of birth, two records with different birthdays are automatically classified as a nonmatch. For accuracy, usually several *blocking passes* are run and a union taken over the matches, since there may be typographical errors in block choices. Deduplication is a special case of record linkage, performed within the same data.

For a more detailed, technical description and history of probabilistic record linkage, refer to Herzog et al. (2007) and Christen (2012). In political science, record linkage is relatively new. Ansolabehere & Hersh (2017) discussed how exact matching between voter data and other sources of administrative data performs surprisingly well, and Enamorado et al. (2018) developed enhanced record linkage open-source software which they tested on national voter files. We use the latter's `fastLink` package; for discussion of its comparative efficacy, see Enamorado et al. (2018). A data dictionary and details about the data we use are in the Supplementary Materials, Section 2.4, and Section 2.5 shows synthetic examples of records' changes between two snapshots.

Evaluating Changes to Voter Files

Matching. Because the snapshots are generated daily, between two snapshots of period $t-1$ and t, the changes we observe are relatively few (median change rate is 1.1%, mean change rate is 5.0%). Excluding exact matches of all variables, we link two consecutive snapshots using first name, last name, street number, postal zip code, and date of birth, employing the Jaro-Winkler string distance and a match threshold of 0.75. The variables are a variation from (1) actual matching criteria implemented by the state of California via their VoteCal system, and (2) the address-date of birth-gender-name combination argued for by Ansolabehere & Hersh (2017). Note that the selection of variables depend on the context and are subject to tuning, and therefore the selection here may not perform well for other jurisdictions. For a performance comparison of the choices of variables, parameters, and string distance metrics, see the Supplementary Materials Sections 3.1 and 3.2.

A record may not be an exact match between two snapshots because (1) it has existed in the previous $t-1$ snapshot but dropped (*dropped record*); (2) because it has not existed in the previous snapshot but newly added in t (*added record*); or (3) because it exists in both snapshots but some field(s) have changed value (*changed record*). In Figure 1, we show the trends of added, dropped, and changed records, as well as changes in key fields such as addresses or party affiliation.

Anomaly Detection. There is no established literature on the dynamics of changes in voter files, save for Pettigrew & Stewart III (2017), who show varying decisions of jurisdictions' on when to remove ineligible voters—i.e., investigating the records dropped. There is no literature that speaks to additions and changes. With no clear prior except the nonstationarity in the data generating process, we need to define the "normal" and "anomalous" volume of changes.

Figure 1 shows anomalies detected by the interquartile range (IQR) method. The IQR method calculates anomalies by calculating the first ($Q1$) and third ($Q3$) quartiles, and isolating the data points outside $[Q1 - x \times \text{IQR}, Q3 + x \times \text{IQR}]$, with the IQR factor x. The usual choice of x is either 1.5 or 3 for moderate or extreme anomaly detection, and our choice the latter. It is a simple first-stage check that can be performed on our audit data as it is intuitive and fast to produce. While alternative anomaly detection methods are possible, we defer this to future research. Prior to applying the IQR method, we detrend the data via seasonal decomposition by piecewise medians.

Results. The first row of Figure 1 shows the records added, dropped, and changed; note that the trends do not always mirror each other. The immediate weeks after elections show very little activity in database updates of any kind, but other dates show varying levels of fluctuations. For instance, for records added, June 29 and December 12 were detected as 'anomalies', while for records deleted, July 26 and December 21 were flagged. Further investigation, coupled with information on from the voter status reason description field, revealed that these were all intentional changes with known administrative causes. For instance, both anomalies for records deleted were outcomes of the National Change of Address (NCOA) processing, which tracks voters who have

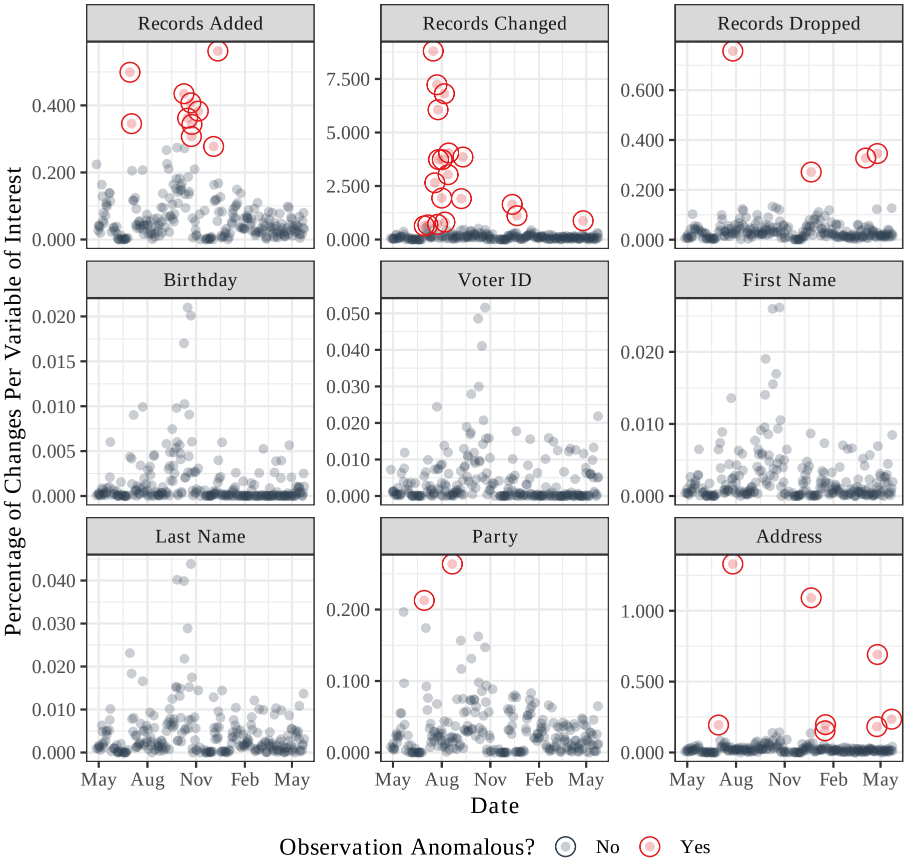


Figure 1: Trend of Changes in Voter Files with Anomalies by the IQR method

moved out of the county—hence rendering the records inactive—using data from the United States Postal Service (USPS). On December 12 the Registrar simultaneously restored the records of voters who were previously inactive but voted on Election Day back to the active registration file.

The remaining panels show field-by-field changes. Party affiliation changes had a local maximum flagged on the primary election registration deadline (May 21). On October 22, the following fields changed significantly: birthdays, first names, last names, and voter IDs, a result of last-minute re-registration by voters immediately before the general registration deadline. For the total number of registrants per snapshot, see Supplementary Materials, Section 2.4. Fortunately, and to the satisfaction of the Registrar, all the anomalies we found were verified by the Registrar as normal administrative activities. No particular degradation of file quality occurred.

Two points of scholarly interest arose from evaluating changes. First, the internal IDs were not always perfectly consistent. The same entity may have, whilst submitting a re-registration, been assigned a new voter ID. Our estimates show that there are more than 13,000 of such cases. This is less than 1% of registrants, but if a researcher plans to use voting history as key covariates, this may skew some intended estimates and must be cautioned against—a single voter would be split into a voter who has not voted after a period, and a voter who newly started to vote.

Second, the generated audit data can be used to study how election officials implement federal and state-level voter registration requirements. For instance, we observed that while the statewide VoteCal system sends NCOA matching data to counties monthly, Orange County in 2018 merged it largely biannually, per their capacity. We were also able to observe the strain that is placed on the county's election administrators by California's system, where registration information can come from the state. Most of the ID changes we observe arose when new information from statewide registration files were sent to county's election management systems, comparative to the ratio of statewide to countywide registration without ID changes. Since internal IDs would ideally stay consistent, this is a possible indication that despite the Registrar's best efforts, some records are not as smoothly merged as might be ideal in the recently-implemented VoteCal database system.

Evaluating Duplicates and Deduplication Efforts

Duplicates are an important indicator in voter file quality (Ansolabehere & Hersh, 2014). Here we quantify changes in incoming and outgoing duplicates, another key indicator of changes in time. When entering a new registration, OCROV clerks may spot a similar voter in the database, but not an exact match. In these cases, they are instructed to keep both, since an accidental deletion can disenfranchise a valid voter, which is worse than keeping a duplicate record. Duplicates are hence an unavoidable part of such administrative data, although the Registrar diligently deduplicates.

The problem is that detecting duplicate voter records are expensive, because it requires extensive human input. In addition, the search procedure for duplicates can be arbitrary. For instance, VoteCal uses driver's license or social security numbers, first name, last name, and date of birth for a "high confidence" duplicate match. Why not also use addresses, emails, phone numbers, and so on? This is difficult to answer a priori—given an unevaluated dataset with no true duplicate status (i.e., a golden rule), how should we detect duplicates so that we may evaluate their trend?

Here we present a simple method to recommend the series of blocking passes for low-cost deduplication with limited resources but high precision. It assesses the efficacy of each block in catching duplicates, and automatically presents them, sorted and with a cumulative number of likely-duplicate record pairs to investigate. Then a user-specified criterion can serve as a cutoff. After we detect duplicates, we can assess how many come and go during the observed period.

Preliminary Blocks. In our first stage, we choose preliminary blocks with combinations of a small number of variables and see how many comparisons each one requires. This initial choice of variables should carry meaningful information on the voter. We choose names, addresses, emails, phone numbers, birthday, and gender. The smallest set of comparison pairs is generated by the block of the first name–birthday–email, generating only 8 pairs, and the largest set by the block of gender–birthday, resulting in 20 million pairs. The full table of preliminary blocks with the reduction ratio, or the efficacy of each block in reducing the comparisons that need to be performed, are in the paper's Supplementary Materials, Section 4.2.

Clearly, some blocks are ineffective, as the gender-birthday block generates potential duplicates almost 14 times greater than the number of total records—this is clearly spurious. Depending on the prior belief about the data, we can exclude blocks that are too noisy to be useful. With Ansolabehere & Hersh (2010)'s 1.5% estimate and OCROV's projection that less than 0.6% of the data are duplicates, we model the distribution of duplicates proportionate to the number of records as $\mathcal{N}(1.1, 0.4)$. Keeping only blocks that generate pairs that are less than 3.5% of the records (a conservative six-sigma bound), we are left with 37 blocking strategies.

Cost Calculus. While we do not have a true duplicate status as a gold standard, we have had an optimal choice of variables when matching snapshots, and we use the same ones to match within blocks, *assuming* that the matches found are true to calculate the rough *potential match rate*. This is shown in Table 1. For example, out of the 8 pairs generated by the first block, 7 were identified as matches and 1 identified as a false positive, with the potential match rate at 87.5%.

We then multiply the number of comparisons to be performed and the *potential non-match rate* to assess how much *false positive cost* each block generates, and scale it from 0 to 100. That

No.	Variables	Number of Comparisons	Match Rate	Cost	Cumulative Matches To-Do
1	First name, Date of birth, Email	8	87.5%	0.00	8
2	First name, Date of birth, Phone	18	100.0%	0.00	23
	...				
13	Last name, First name, Address (full)	7,325	100.0%	0.00	7,781
14	Last name, First name, Address (part)	7,690	100.0%	0.00	8,132
15	Gender, Date of birth, Phone	215	97.7%	0.02	8,322
	...				
21	Date of birth, Address (full)	3,462	99.8%	0.04	12,123
22	Date of birth, Address (part)	4,691	99.8%	0.04	12,784
23	First name, Address (full)	9,454	99.7%	0.14	14,746
24	Last name, Gender, Email	734	90.3%	0.33	15,413
25	Last name, First name, Phone	764	88.4%	0.41	15,533
	...				
36	Gender, Phone	25,344	58.4%	48.38	86,329
37	First name, Address (part)	41,618	47.7%	100.00	118,109

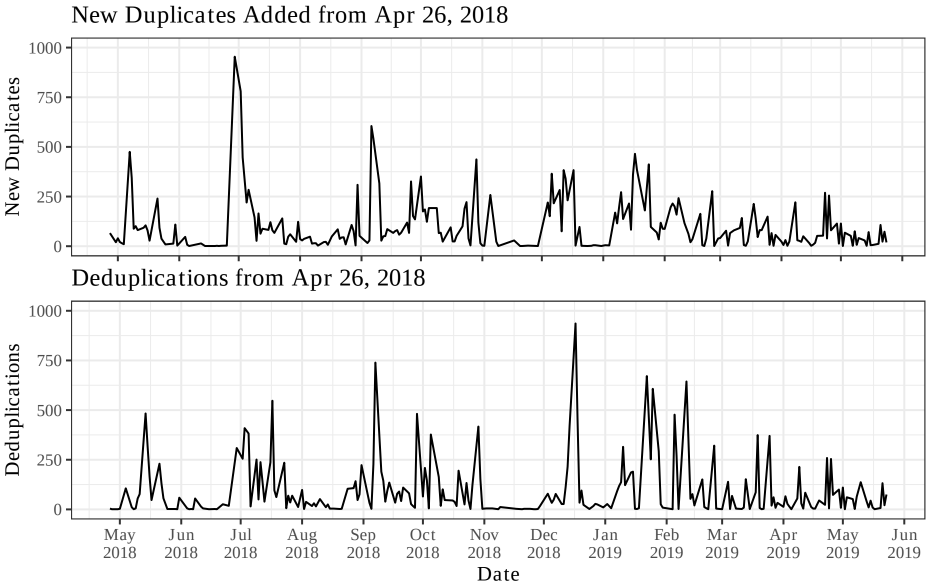
Table 1: Cost Comparison of Blocks, April 26 (Full Table in Appendix)

is, we aim to minimize the erroneous decision to classify two separate voters as a single voter. Table 1 shows the blocks aligned from low- to high-cost. The cost may be high because (1) there are too many potential comparisons to be performed, or (2) out of the comparisons, too many are non-matches, both contributing to lowering precision. The last column of Table 1 shows the *cumulative matches to do* that corrects for overlap in potential duplicates produced by each block.

Results. This yardstick can be compared with a user-specified threshold that uses prior information on the voter file and available resources. If the prior is that less than 8,000 pairs will be duplicates, we can choose to use the top 13 blocks, resulting in 7,781 clerical reviews to perform. While this is practically useful for a Registrar with a limited budget, the Table itself shows that the match rate is extremely high for blocks in a few subsequent rows as well, implying the threshold must be raised. For efficiency, we choose the first 23 blocks—again, this is not an exact measure of duplicates, but an excellent approximate while avoiding false positives.

We can then use the chosen blocks to find potential duplicates across all our current and future snapshots, and monitor the quantities of incoming duplicates and deduplication efforts, just as we have monitored changes in Figure 1. Figure 2 shows the trend of incoming and outgoing duplicates. Upon scrutiny, we found that most duplicates came from state-driven changes, relative to the statewide vs. countywide registration in non-duplicate new registrations. Again, this could be signaling the strain from multiple governmental agencies sending data per the National Voter

Registration Act of 1993, something yet to be explored. This result shows that audit data our method generate may provide useful information for further study by researchers, as they help shed light on how election officials grapple with administrative and technological changes.


New Duplicates Added from Apr 26, 2018

Deduplications from Apr 26, 2018

Figure 2: Trend of Incoming and Outgoing Duplicates

Discussion

We developed two complementary methods that help evaluate and improve the quality of a voter registration database as it changes over time, using 2018-2020 data from Orange County, California. The first method generates audit data of periodic changes to the database as a time-series, which can undergo anomaly detection to check for internal and external unwanted changes and the deterioration in database quality. The second method presents a set of potential duplicates using an automated procedure that is efficient and which minimizes false positives; it allows us to track new duplicates and deduplication efforts over time, in addition to identifying the sources of new duplicates.

Our methods focus on assessing the internal validity of voter data, though adding other methods that can evaluate their external validity in the future will be important extensions of our research. Both of the methods we present yield data that can be further scrutinized—for election officials, it enables forensics and affirms their integrity, and for scholars, it presents interesting new data that can be used to study how election administrators implement voter registration requirements.

References

Alvarez, R. M., Ansolabehere, S., & Stewart, C. (2005). Studying elections: Data quality and pitfalls in measuring the effects of voting technologies. *Policy Studies Journal*, *33*(1), 15-24.

Alvarez, R. M., Jonas, J., Winkler, W. E., & Wright, R. N. (2009). *Interstate voter registration database matching: The Oregon-Washington 2008 pilot project*. Proceedings of Workshop on Trustworthy Elections.

Ansolabehere, S., & Hersh, E. (2010). The quality of voter registration records: A state-by-state analysis. *Report, Caltech/MIT Voting Technology Project*.

Ansolabehere, S., & Hersh, E. (2014). Voter registration: the process and quality of lists. *The measure of American elections*, 61–90.

Ansolabehere, S., & Hersh, E. (2017). ADGN: An algorithm for record linkage using address, date of birth, gender, and name. *Statistics and Public Policy*, *4*, 1-10.

Christen, P. (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media.

Enamorado, T., Fifield, B., & Imai, K. (2018). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 1–19.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*, 1183-1210.

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer.

Pettigrew, S., & Stewart III, C. (2017). Moved out, moved on: Assessing the effectiveness of voter registration list maintenance. *MIT Political Science Department Research Paper No. 2018-1*. doi: 10.2139/ssrn.3044810

Sanger, D. E. (2018). *The perfect weapon: War, sabotage, and fear in the cyber age.* Crown.

Evaluating the Quality of Changes in Voter Registration Databases: Supplementary Materials

1 Acknowlegements

We thank the John Randolph and Dora Haynes Foundation for providing funding to support the research reported in this paper. Schneider thanks the Caltech Summer Undergraduate Research Fellowship program for supporting his work on this project. We thank Neal Kelley and his team at the Orange County Registrar of Voters (OCROV) office for their partnership on this project. We especially thank Justin Berardino of the OCROV for answering our many questions, and facilitating the transfer of the data we use in this paper.

2 Data

2.1 Data Overview

In this paper, we investigate 252 unique daily snapshots of the Orange County Voter Registration dataset, beginning April 26, 2018, and ending May 24, 2019. Altogether, they cover 89% of business days (weekdays). Each snapshot consists of roughly 1.5 million voters. We continue to receive daily snapshots of the OC dataset in the 2020 cycle.

2.2 Why Orange County?

Orange County (California) is a large and diverse county in Southern California. Located south of Los Angeles and north of San Diego, Orange County is home to a wide array of different business, colleges and universities, and of course, Disneyland. The county currently has a total population of almost 3.2 million residents, and in the 2016 presidential election, Orange County had just over 2 million voting-eligible citizens, with approximately 1.5 million registered voters California Secretary of State (2016). In that same election, 1.2 million of those registered voters participated (80.7% of registered voters). Orange County's population is also diverse, as the U.S. Census Bureau's most recent estimates show that 72% of the county's population is White, 21% Asian, 2% African-American, and 3.5% two or more races. The Census Bureau's recent data estimates that 34% of the Orange County's population is Hispanic or Latino United States Census Bureau (2017). Thus, one reason we focus on Orange County for this study is that it is one of the largest and most diverse election jurisdictions in the United States.

Secondly, Orange County is widely viewed as an innovator in the administration of elections. The County's Registrar of Voters, Neal Kelley, participates widely in state and national professional organizations, and is has been recognized for his innovative administrative practices. Under his administration, Orange County has developed many administrative processes and tools that are viewed as best practices for election administration. These innovations include, for example, building transparency by webcasting in real time virtually all aspects of the process of administering an election, or more recently, pilot testing risk-limiting audits.

2.3 Data Availability

Upon publication, all of the code necessary to produce the analyses reported in our paper will be available on the GitHub repository `REDACTED`, along with an example dataset with synthetic voter information. Due to the confidential nature of the voter registration data, and our data access agreement with OCROV, we cannot share or post publicly the data used in this study. Researchers who want to use these data can request access from the Orange County Registrar of Voters.

2.4 Data Dictionary

The voter file "snapshots" that we have received from the OCROV contain the fields described below. The number in parentheses describe the number of unique values for each field,[1] based on the snapshot of May 21, 2018, the registration deadline for the June 2018 primaries. The snapshot consists of 1,478,541 observations.

Here we provide a data dictionary and the number of unique values in each of the sixty-two data fields.[2] Many of the variables are created internally by the Orange County Registrar of Voters for their usage; our interest is mostly limited to variables that contain direct inputs from the voters. These variables of interest are listed in the Appendix in Table 2 with summary statistics.[3] Although the Registrar assigns each voter with a unique ID (`lVoterUniqueID`) that is not duplicated in any of the daily snapshots, not all voters are distinct entities.

In Orange County, the voter registration forms ask the voter for both the California Driver's License number (or a California Identification card number) and the last four digits of the Social Security Number (SSN) Orange County Registrar of Voters (2018b). However, these are not strictly

[1] The numbers are based on raw text, so that for instance, "MISS" and "Miss" are counted as distinct values.

[2] Note that the canonical text cleaning and standardizing precedes both the calculations of number of unique entries and the occurrence of the most frequent entries, such as stripping the string of non-alphanumeric entries, trimming white-spaces, and case normalizing, except for email addresses, which may be case sensitive and in which certain punctuation creates meaningful differences.

[3] We exclude mailing addresses due to the fact that it usually overlaps with physical, residential address. We also excluded reported place of birth as it seems to frequently be misreported, and the reported place of birth changes frequently in the data.

required. If neither of them can be provided, a voter may be assigned a unique ID number solely for registration purposes (Orange County Registrar of Voters, 2018a). Despite these seemingly unique identifiers, duplicates still can be found in the database. Indeed, deduplication based on exact matching on these identifiers—the most basic of deduplication efforts—is already performed by the OCROV.

- “lVoterUniqueID” (1,478,541): Interally assigned voter identification number.
- “sAffNumber” (1,478,540): An identifier of the voter registration affidavit.
- “szStateVoterID" (1): The voter identification number assigned by the Secretary of State’s Office to the record.
- “sVoterTitle” (10): Title (e.g., “Dr.”, “Mrs.”) provided by the voter.
- “szNameLast” (188,734): Last name.
- “szNameFirst” (89,985): First name.
- “szNameMiddle” (52,085): Middle name.
- “sNameSuffix” (23): Name suffix.
- “sGender” (3): Gender.
- “szSitusAddress” (787,043): Address.
- “szSitusCity” (48): City.
- “sSitusState” (1): State.
- “sSitusZip” (94): Zip Code.
- “sHouseNum” (30,269): House number.
- “sUnitAbbr” (20): House unit abbreviation.
- “sUnitNum” (14,780): House unit number.
- “szStreetName” (17,437): Street name.
- “sStreetSuffix” (95): Street suffix.
- “sPreDir” (9): Direction prefix.
- “sPostDir” (5): Direction suffix.
- “szMailAddress1” (807,272): Mailing address (street address).
- “szMailAddress2” (22,249): Mailing address (city, state, and zip code).
- “szMailAddress3” (2,271): Mailing address (overseas voters’ street address).
- “szMailAddress4” (195): Mailing address (overseas voters’ country of residence).
- “szMailZip” (13,425): Mailing Zip Code.
- “szPhone” (706,711): Telephone number.
- “szEmailAddress” (452,610): Email address.
- “dtBirthDate” (30,468): Date of birth.
- “sBirthPlace” (30,468): Place of birth.
- “dtRegDate” (15,762): Registration record date.
- “dtOrigRegDate” (16,477): Original registration date.

- “dtLastUpdate_dt” (6,984): Update of record.
- “sStatusCode” (1): Status of record.
- “szStatusReasonDesc” (110): Description of record status.
- “sUserCode1” (7,370): (Unknown)
- “sUserCode2” (13): (Unknown)
- “iDuplicateIDFlag” (4): Potential duplicate ID flag.
- “szLanguageName” (1): Language.
- “szPartyName” (46): Party registration.
- “szAVStatusAbbr” (12): Absentee status abbreviation.
- “szAVStatusDesc” (12): Absentee status description.
- “szPrecinctName” (53): Precinct name.
- “sPrecinctID” (1,487): Precinct ID.
- “sPrecinctPortion” (8): Precinct portion.
- “sDistrictID_0” (1): Geographic district identifier (0: County).
- “iSubDistrict_0” (1): Geographic district (0: County).
- “szDistrictName_0” (1): Geographic district name (0: County).
- “sDistrictID_1” (7): Geographic district identifier (1: Congressional district).
- “iSubDistrict_1” (1): Geographic district (1: Congressional district).
- “szDistrictName_1” (7): Geographic district name (1: Congressional district).
- “sDistrictID_2” (5): Geographic district identifier (2: Senate district).
- “iSubDistrict_2” (1): Geographic district (2: Senate district).
- “szDistrictName_2” (5): Geographic district name (2: Senate district).
- “sDistrictID_3” (7): Geographic district identifier (3: Assembly district).
- “iSubDistrict_3” (1): Geographic district (3: Assembly district).
- “szDistrictName_3” (7): Geographic district name (3: Assembly district).
- “sDistrictID_4” (5): Geographic district identifier (4: Supervisorial district).
- “iSubDistrict_4” (1): Geographic district (4: Supervisorial district).
- “szDistrictName_4” (5): Geographic district name (4: Supervisorial district).
- “sDistrictID_5” (35): Geographic district identifier (5: City council ward division).
- “iSubDistrict_5” (9): Geographic district (5: City council ward division).
- “szDistrictName_5” (68): Geographic district name (5: City council ward division).

2.5 Hypothetical Changes to the Database

Table 1 shows synthetic examples of changes in the voter file. They can also represent examples of duplicates in the file.

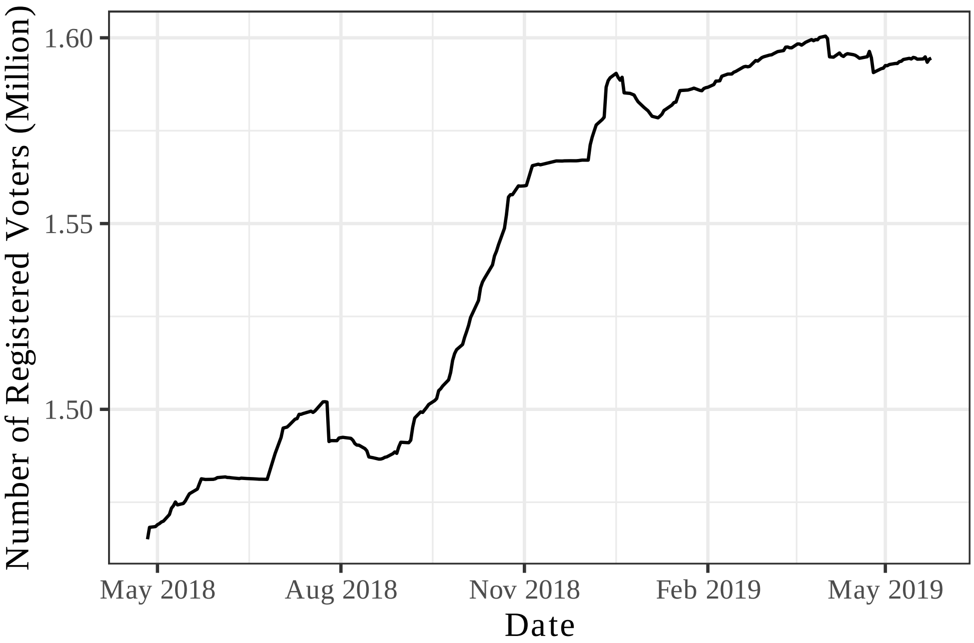


Figure 1: Number of Records Per Day

Name			Address		Birth Date	Contact	
First	Middle	Last	Street Address	City		Phone	Email
Steven	B	Smith	110 S East Ave	Brea	04/26/1980	714-765-3300	N/A
Steven		Smith	110 S East Ave	Brea	04/26/1980	714-765-3300	smith@ex
Isidor		Agnes	99 6th St #72	Tustin	07/13/1960	N/A	N/A
Jsidor		Agne	99 6th St #72	Tustin	07/13/1960	714-205-8583	N/A
Anna	Clara	Zhang	203 Coast Ln	Tustin	12/01/1950	N/A	acz@ex
Anna	C	Zhang	101 Sunny Blvd	Brea	12/10/1950	N/A	acz@ex

Table 1: Synthetic Examples of Changes in Voter Files

2.6 Descriptive Statistics

Figure 1 show the total number of observations in the voter registration database by date. As can be seen, the daily snapshots were generated on business days (weekdays). There are a few missing snapshots—while the Orange County Registrar of Voters have made incredible contributions by providing us with daily snapshots, when they were busy, we were unable to obtain some snapshots.

In addition, as aforementioned, Table 2 shows the data summary for some important user-entered variables. This shows how data-intensive each field is, showing the amount of missing data for the important fields, and the number of unique and most frequent entries. For instance, the name suffix has too much missing data and too few unique entries to be very informative. Political party, although an important variable, is likewise not informative for matching.

Table 2: Data Summary by Field of May 21 Snapshot

Category	Field	Number of Unique Entries	Number of Most Freq. Entry	Number Missing	Examples
Name	First	89,984	21,481	78	Jane
	Middle	51,609	83,035	406,428	E
	Last	188,734	26,385	0	Doe
	Title (Name Prefix)	5	466,043	488,123	Ms.
	Name Suffix	18	16,430	1,452,055	Jr.
Address	Street Address	786,224	93	0	1300 S Grand Ave Unit 101
	City	48	140,081	0	Santa Ana
	Zip Code	94	40,128	0	92705
Date of Birth		30,467	124	23	March 11, 1989
Place of Birth		319	678,187	60,999	CA
Gender		3	2,274	1,474,151	F
Political Party		46	540,859	0	No Party Preference
Contact	Phone	706,710	9,035	663,105	(714) 567-7600
	Email	452,609	382	1,018,894	jane@roc.ocgov.com

3 Parameter and Variable Selection in Record Linkage

A recap of the probabilistic record linkage framework, which forms the basis of our analysis, is in Figure 2. The two density distributions show match probability by the latent status of a true match. If the match is a "true negative," i.e, the entities are not the same voter, the match probability is likely lower than when the match is a "true positive." However, due to chance, some fields such as names or address may coincide, resulting in an overlapping region. A researcher typically decides upon a lower and upper cutoff of the match probability to classify the record pairs into nonmatches, matches, and those that must be clerically reviewed. Note that for the final composite match probability, we have to calculate each fields' agreement levels and weight it using its frequency

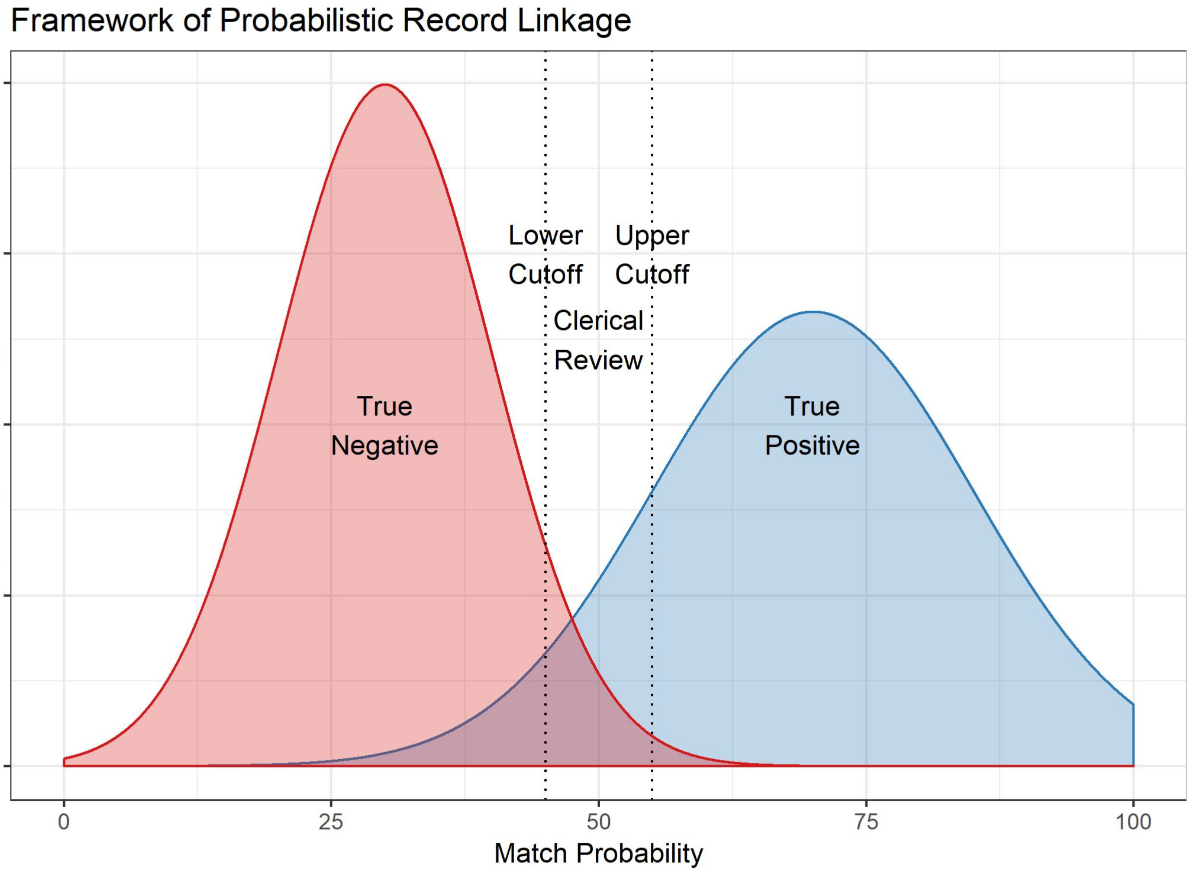


Figure 2: The Framework of Probabilistic Record Linkage

distribution.

3.1 String Distance Metrics and Threshold

In this Section we briefly explore how we chose the parameters in record linkage. As we have aforementioned in the main text, we use `R` and its CRAN package `fastLink`. While there are many different options in `fastLink`, the following are major parameters of choice: the choice of the string distance metric (`stringdist.method`), and the cutoff threshold that declares a match (`threshold.match`). The first determines the spectrum of the agreement between two strings. The second determines the lower cutoff for a match classification—that is, in Figure 2, we only use a single cutoff for simplicity, not leaving any records for clerical review.

The default values of each are respectively the Jaro-Winkler string distance metric and a threshold of 0.85. We test out the following combination of string metric-threshold parameters:

$$c(\text{Jaro-Winkler, Levenshtein distance}) \times c(0.70, 0.75, 0.80, 0.85, 0.90, 0.95)$$

While we do not have a gold standard—i.e., true match status—when matching between snapshots, we have a good alternative for it, which is the internal ID assigned within the OCROV. Assuming that it is the true match status, we can employ the following canonical performance measurements in record linkage: pairwise precision, pairwise recall, and F1 score. For details on the string distance measures and the performance metrics, refer to Christen (2012).

The followings are the performance matrices for the twelve parameter combinations, using the variables mentioned in the main text.

No.	String Metric	Threshold	Precision	Recall	F1
1	Jaro-Winkler	0.70	0.9433	0.9939	0.9657
2	Jaro-Winkler	0.75	0.9364	0.9970	0.9672
3	Jaro-Winkler	0.80	0.9343	0.9970	0.9659
4	Jaro-Winkler	0.85	0.9429	0.9873	0.9600
5	Jaro-Winkler	0.90	0.9360	0.9972	0.9671
6	Jaro-Winkler	0.95	0.9377	0.9937	0.9659
7	Levenshtein	0.70	0.9386	0.9793	0.9514
8	Levenshtein	0.75	0.9356	0.9623	0.9487
9	Levenshtein	0.80	0.9306	0.9689	0.9427
10	Levenshtein	0.85	0.9342	0.9970	0.9659
11	Levenshtein	0.90	0.9397	0.9873	0.9671
12	Levenshtein	0.95	0.9256	0.9524	0.9290

Table 3: Performance Evaluation for String Distance Metric and Threshold Choices

Note that because the internal ID may be inconsistent, some of the matches that are classified as false are true matches. There are no cases vice versa to our knowledge, i.e., cases where two people share the same internal voter ID. Hence pairwise precision is slightly undervalued, and as a result the F1 score. We still use F1 score as our final metric for tuning as it is a harmonic mean between precision and recall.

In the grid that we explored, it seems to be the case that the Jaro-Winkler string metric combined with a threshold value of 0.75 works best. Note that the threshold value of choice is lower than the default value in Enamorado et al. (2018). Also note that while 0.75 works best when the string distance metric is fixed to Jaro-Winkler, 0.90 works best when the metric is Levenshtein distance. When the threshold value is fixed at 0.85, Levenshtein distance performs better. While we have chosen optimal parameters, this also is a cautionary tale in applying record linkage in other datasets and other domains.

3.2 Variable Selection

Another choice that the researcher should make when employing probabilistic record linkage is to choose which variables to perform the matching on. This depends substantially on the dataset's existing variables and the dataset's size. Our first intuition was the 7th combination of variables: first name, last name, date of birth, street number, and zip code. We test the performance for adding or deleting variables from this combination, using the tuned parameters from above.

No.	Variables	Precision	Recall	F1
1	First name, middle name, last name, date of birth, street number, zip code	0.9648	0.6237	0.7529
2	First name, last name, date of birth, street number, street name, zip code	0.9383	0.8190	0.8744
3	First name, last name, date of birth, street number, house number, zip code	0.9411	0.1864	0.3203
4	First name, last name, date of birth, street number, street name, house number, zip code	0.9366	0.1884	0.3206
5	First name, last name, date of birth, full street address, zip code	0.9354	0.8493	0.8874
6	First name, last name, date of birth, gender, street number, zip code	0.9432	0.9106	0.9264
7	First name, last name, date of birth, street number, zip code	0.9364	0.9970	0.9672
8	First name, date of birth, street number, zip code	0.9219	0.9843	0.9593
9	Last name, date of birth, street number, zip code	0.9313	0.9832	0.9605
10	First name, last name, street number, zip code	0.9334	0.9595	0.9516
11	First name, last name, date of birth, zip code	0.9361	0.9682	0.9542
12	First name, last name, date of birth, street number	0.9361	0.9684	0.9654

Table 4: Performance Evaluation for Variable Choices

The initial combination of choice seems to be indeed most optimal in terms of the F1 score. The next-best choice seems to be using only the street number. Note that adding variables seem to cause much more damage by creating false negatives and decreasing the recall. Precision is relatively robust. Regardless, hence our choice of variables used to match snapshots is the seventh set of variables.

4 Duplication Detection

4.1 Setup

The cheapest duplicate detection methods are often exact matches that are *rule-based* Hernandez & Stolfo (1998), i.e., a researcher defines specifically what a match is—for instance, a match may be declared when first name, last name, date of birth, and residing city are exact matches. With fuzzy matches, we can reduce computational costs by *blocking* to reduce comparison pairs as aforementioned. However, the choice of rules or blocks both requires extensive "domain-specific expertise" as the literature puts it, which makes it difficult to automate the selection, and has room for arbitrary choices. The procedure here describes an automated measure to lessen this problem, but the initial choice of variables do need some knowledge about the data, as aforementioned in the main text.

We use a combination of two or three of the following variables:

- Last name
- First name

- Date of birth
- Residential address (part): Street number, street name, zip code
- Phone number
- Email address

In addition, we give the following variations: for all combinations with first names, generate blocks that substitute first name for gender, and for all combinations with address (part), generate blocks that substitute street number and street name for a full single string of street address, including directions and unit numbers. The rationale is that the OCROV data contains very little information in the existing gender field, so that gender is largely inferred from first names, given prefixes, and a few scattered entries of exiting ‘sGender’. Hence including both made little sense. The latter is motivated by the fact that apartment numbers are often missing and street directions as well (e.g. North, South, East, West). The full street address contains the parts of the addresses. Moreover, we include two blocks of single variables: address (part) and address (full). On the other hand, we leave out gender-last name block, because the blocks were too big and crashed the available computation resources when computed in a 320G memory.

This leaves us with seventy-one blocks to be tested. Generation of blocks were performed with CRAN package `RecordLinkage`, as at the beginning of the project, `fastLink` did not have the means to preprocessing matches via blocking. Table 5 shows the blocks aligned by reduction ratio. It also displays the key distribution statistics of the block sizes (the minimum is always 1), the number of blocks (i.e., unique values), and the number of non-missing occurrences.

Table 5: Cost Comparison for Blocks: Pre-Matching, April 26 Snapshot

				Block Size Distribution					
No.	Variables	Non-missing Obs.	Number of Blocks	Q1	Q2	Q3	Max	Number of Comparisons	Reduction Ratio (%)
1	First name, Date of birth, Email	30.7%	449,848	1	1	1	2	8	99.9999999996
2	First name, Date of birth, Phone	54.4%	797,257	1	1	1	2	18	99.9999999992
3	First name, Phone, Email	22.3%	326,703	1	1	1	2	19	99.9999999991
4	Gender, Date of birth, Email	28.8%	421,535	1	1	1	2	23	99.9999999989
5	Date of birth, Phone, Email	22.3%	326,703	1	1	1	2	26	99.9999999988
6	Last name, Date of birth, Email	30.7%	449,839	1	1	1	2	31	99.9999999986
7	First name, Date of birth, Address (full)	100.0%	1,464,891	1	1	1	2	42	99.9999999980
8	First name, Address (full), Email	30.7%	449,806	1	1	1	2	55	99.9999999974
9	First name, Date of birth, Address (part)	99.9%	1,463,820	1	1	1	2	56	99.9999999974
10	Date of birth, Address (full), Email	30.7%	449,814	1	1	1	2	56	99.9999999974
11	Date of birth, Address (part), Email	30.7%	449,324	1	1	1	2	57	99.9999999973
12	Last name, First name, Email	30.7%	449,803	1	1	1	2	58	99.9999999973
13	First name, Address (part), Email	30.7%	449,314	1	1	1	2	58	99.9999999973
14	Date of birth, Email	30.7%	449,802	1	1	1	2	68	99.9999999968
15	First name, Email	30.7%	449,777	1	1	1	2	84	99.9999999961
16	Gender, Date of birth, Phone	51.1%	747,693	1	1	1	2	215	99.9999999900
17	Last name, First name, Date of birth	100.0%	1,464,693	1	1	1	3	241	99.9999999888
18	Last name, Date of birth, Phone	54.4%	797,014	1	1	1	3	291	99.9999999864
19	Date of birth, Address (full), Phone	54.4%	796,952	1	1	1	3	353	99.9999999836
20	Date of birth, Address (part), Phone	54.4%	796,366	1	1	1	3	363	99.9999999831
21	Gender, Phone, Email	20.9%	305,835	1	1	1	4	409	99.9999999809
22	Date of birth, Phone	54.4%	796,882	1	1	1	3	423	99.9999999803
23	Last name, Gender, Email	28.8%	420,860	1	1	1	4	734	99.9999999658
24	Last name, First name, Phone	54.4%	796,522	1	1	1	3	764	99.9999999644

25	First name, Address (full), Phone	54.4%	796,512	1	1	1	3	775	99.9999999639
26	First name, Address (part), Phone	54.4%	795,919	1	1	1	3	792	99.9999999631
27	First name, Phone	54.4%	796,315	1	1	1	3	972	99.9999999547
28	Gender, Address (full), Email	28.8%	420,590	1	1	1	7	1,035	99.9999999518
29	Gender, Address (part), Email	28.7%	420,083	1	1	1	7	1,073	99.9999999500
30	Gender, Email	28.8%	420,377	1	1	1	7	1,269	99.9999999409
31	Last name, Phone, Email	22.3%	324,807	1	1	1	4	2,004	99.9999999066
32	Gender, Date of birth, Address (full)	94.0%	1,375,275	1	1	1	4	2,024	99.9999999057
33	Address (full), Phone, Email	22.3%	324,368	1	1	1	4	2,488	99.9999998841
34	Address (part), Phone, Email	22.3%	323,943	1	1	1	4	2,543	99.9999998815
35	Gender, Date of birth, Address (part)	93.9%	1,373,733	1	1	1	4	2,565	99.9999998805
36	Phone, Email	22.3%	324,214	1	1	1	4	2,670	99.9999998756
37	Last name, Date of birth, Address (full)	100.0%	1,462,355	1	1	1	4	2,703	99.9999998741
38	Last name, Date of birth, Address (part)	99.9%	1,461,213	1	1	1	4	2,790	99.9999998700
39	Date of birth, Address (full)	100.0%	1,461,602	1	1	1	4	3,462	99.9999998387
40	Last name, Address (full), Email	30.7%	445,404	1	1	1	4	4,673	99.9999997823
41	Date of birth, Address (part)	99.9%	1,459,367	1	1	1	4	4,691	99.9999997814
42	Last name, Address (part), Email	30.7%	444,829	1	1	1	4	4,765	99.9999997780
43	Last name, Email	30.7%	445,117	1	1	1	6	4,995	99.9999997673
44	Address (full), Email	30.7%	443,984	1	1	1	10	6,275	99.9999997076
45	Address (part), Email	30.7%	443,355	1	1	1	10	6,434	99.9999997002
46	Last name, First name, Address (full)	100.0%	1,457,666	1	1	1	3	7,325	99.9999996587
47	Last name, First name, Address (part)	99.9%	1,456,258	1	1	1	4	7,690	99.9999996417
48	First name, Address (full)	100.0%	1,455,598	1	1	1	4	9,454	99.9999995595
49	Last name, Gender, Phone	51.1%	733,684	1	1	1	6	15,292	99.9999992875
50	Gender, Address (full), Phone	51.1%	730,463	1	1	1	18	19,175	99.9999991066
51	Gender, Address (part), Phone	51.0%	729,561	1	1	1	18	19,593	99.9999990871
52	Last name, Gender, Date of birth	94.0%	1,358,093	1	1	1	16	21,173	99.9999990135
53	Gender, Phone	51.1%	724,976	1	1	1	18	25,344	99.9999988192
54	First name, Address (part)	99.9%	1,434,127	1	1	1	17	41,618	99.9999980610
55	Last name, Date of birth	100.0%	1,424,977	1	1	1	20	49,456	99.9999976958
56	Last name, Address (full), Phone	54.4%	728,639	1	1	1	8	75,722	99.9999964720
57	Last name, Address (part), Phone	54.4%	727,208	1	1	1	8	76,695	99.9999964267
58	Last name, Phone	54.4%	723,423	1	1	1	8	82,495	99.9999961565
59	Address (full), Phone	54.4%	710,824	1	1	1	30	98,511	99.9999954102
60	Address (part), Phone	54.4%	708,911	1	1	1	31	100,205	99.9999953313
61	First name, Date of birth	100.0%	1,348,674	1	1	1	8	138,931	99.9999935270
62	Last name, Gender, Address (full)	94.0%	1,220,017	1	1	1	10	181,918	99.9999915242
63	Last name, Gender, Address (part)	93.9%	1,209,649	1	1	1	42	203,839	99.9999905029
64	Gender, Address (full)	94.0%	1,100,303	1	1	1	65	348,869	99.9999837457
65	Last name, Address (full)	100.0%	1,010,761	1	1	2	11	619,849	99.9999711204
66	Last name, Address (part)	99.9%	993,131	1	1	2	95	682,501	99.9999682014
67	Address (full)	100.0%	783,022	1	2	2	94	1,049,036	99.9999511240
68	Last name, First name	100.0%	1,129,285	1	1	1	278	2,598,749	99.9998789209
69	Gender, Address (part)	93.9%	901,372	1	1	2	418	5,779,992	99.9997307026
70	Address (part)	99.9%	596,116	1	2	3	691	13,091,823	99.9993900349
71	Gender, Date of birth	94.0%	58,985	13	25	33	65	20,016,417	99.9990674090

4.2 Assessing Blocks

We then calculate the potential match rate, false positive costs, and finally, after aligning the blocks by their costs, calculate the cumulative matches to do, as explained. The following Table 6 is a full version of the Table 2 in the main text.

Table 6: Cost Comparison for Blocks: Pre-Matching, April 26 Snapshot

No.	Variables	Number of Comparisons	Match Rate	Cost	Cumulative Matches To-Do
1	First name, Date of birth, Email	8	87.5%	0.00	8
2	First name, Date of birth, Phone	18	100.0%	0.00	23
3	First name, Phone, Email	19	100.0%	0.00	39

4	Gender, Date of birth, Email	23	95.7%	0.00	54
5	First name, Date of birth, Address (full)	42	100.0%	0.00	91
6	First name, Address (full), Email	55	100.0%	0.00	128
7	First name, Date of birth, Address (part)	56	100.0%	0.00	142
8	Last name, First name, Email	58	98.3%	0.00	156
9	First name, Address (part), Email	58	100.0%	0.00	159
10	Last name, First name, Date of birth	241	100.0%	0.00	366
11	First name, Address (full), Phone	775	99.9%	0.00	1,122
12	First name, Address (part), Phone	792	99.9%	0.00	1,139
13	Last name, First name, Address (full)	7,325	100.0%	0.00	7,781
14	Last name, First name, Address (part)	7,690	100.0%	0.00	8,132
15	Gender, Date of birth, Phone	215	97.7%	0.02	8,322
16	Gender, Date of birth, Address (full)	2,024	99.8%	0.02	10,128
17	Gender, Date of birth, Address (part)	2,565	99.8%	0.02	10,653
18	Last name, Date of birth, Address (full)	2,703	99.9%	0.02	11,668
19	Last name, Date of birth, Address (part)	2,790	99.9%	0.02	11,695
20	First name, Email	84	90.5%	0.04	11,701
21	Date of birth, Address (full)	3,462	99.8%	0.04	12,123
22	Date of birth, Address (part)	4,691	99.8%	0.04	12,784
23	First name, Address (full)	9,454	99.7%	0.14	14,746
24	Last name, Gender, Email	734	90.3%	0.33	15,413
25	Last name, First name, Phone	764	88.4%	0.41	15,533
26	First name, Phone	972	85.8%	0.63	15,578
27	Gender, Phone, Email	409	59.9%	0.75	15,738
28	Gender, Address (full), Email	1,035	66.8%	1.58	15,975
29	Gender, Address (part), Email	1,073	66.3%	1.66	15,991
30	Gender, Email	1,269	57.3%	2.49	16,085
31	Last name, Gender, Phone	15,292	87.5%	8.80	30,245
32	Last name, Gender, Date of birth	21,173	84.9%	14.65	49,442
33	Gender, Address (full), Phone	19,175	72.1%	24.55	55,284
34	Gender, Address (part), Phone	19,593	71.7%	25.46	55,492
35	Last name, Date of birth	49,456	84.5%	35.21	82,730
36	Gender, Phone	25,344	58.4%	48.38	86,329
37	First name, Address (part)	41,618	47.7%	100.00	118,109

References

California Secretary of State. (2016). *Voter registration statistics by county: Report of registration as of October 24, 2016.* Retrieved from `http://elections.cdn.sos.ca.gov/sov/2016-general/sov/02-voter-reg-stats-by-county.pdf`

Christen, P. (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection.* Springer Science & Business Media.

Enamorado, T., Fifield, B., & Imai, K. (2018). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 1–19.

Hernandez, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, *2*, 9-37.

Orange County Registrar of Voters. (2018a). *Citizenship & voting: Register to vote.* Retrieved from `http://www.ocgov.com/residents/voting/register`

Orange County Registrar of Voters. (2018b). *Register to vote, or change your name, address or party.* Retrieved from `https://www.ocvote.com/registration/register-to-vote/`

United States Census Bureau. (2017). *Quickfacts: Orange County, California.* Retrieved from `https://www.census.gov/quickfacts/fact/table/orangecountycalifornia/PST045217`