Voter Removal from Registration List
Based on Name Matching is Unreliable

Ted Selker, Alexandre Buer
Voting Technology Project - MIT Media Laboratory

7 April, 2006

Abstract

The voter registration list is the information backbone for the administration of elections. Keeping it up-to-date is a difficult task that can expose officials to accusations of voter disenfranchisement. We review here some of the problems that affected Florida elections, explore some solutions proposed with the Help America Vote Act of 2002 in regards to voter registration maintenance, and illustrate these with an experiment on the actual voter rolls from Florida.

The voter registration list (VR) is the information backbone for the administration of elections. Kept in computerized form on database systems, election officials use it to estimate the size of the electoral population, and in turn the budget, staffing, office space, and quantities of printed materials necessary to conduct elections. Thus there is a strong public interest in keeping the VR current and accurate. This process is usually called voter registration list maintenance. This maintenance includes adding newly registered citizens or citizen who moved into the jurisdiction, correcting records with name spelling, date of birth, or address mistakes, updating addresses of voters who moved within the same jurisdiction that the VR covers, removing duplicate records (i.e., multiple records that point to a unique voter), removing voters who moved to addresses outside of that jurisdiction, removing voters who died, and removing persons that are barred from voting by state law (e.g., convicted felons in some states). The process of removing persons from voter registration lists is often called a purge.

The administration of elections is not uniform in the U.S. In many states each county administers their own elections, with a large degree of independence on the implementation details. Although centralization efforts at the state level started before the general election of 2000, many counties still maintained their own list, using their own system and staff. In reaction to the lack of uniformity and problems in the election system revealed for the 2000 general election, Congress passed the Help America Vote Act of 2002 [HAVA]. The most important change for the voter registration process is the move towards a “single, uniform, official, centralized, interactive computerized statewide voter registration list defined, maintained, and administered at the State level.” [HAVA 303(a)(1)].

The Act also details that the voter registration list must include either the driving license number or the last four digit of the social security number. A unique universal identifier (a number guaranteed to be unique to only one person and to apply to all eligible voters) managed by a central authority is a proven solution to manage a large database. It allows for the comparison of records between different databases, given that they use the same identifier to manage records. The social security number is one example, designed by the Social Security Administration to help it identify persons and their benefits based on their work and participation in the system. Unfortunately, many non-governmental institutions also use this key identifier as a token to prove one's identity and to provide authentication. This has led to the perverse situation where HAVA mandates that the SSN not be used in the VR system. The laudable objective of reducing the vulnerability of the system results in the social security number being considered confidential, which is the main reason behind the decision to allow only the use of the last four digits of the social security number in voter registration systems. This makes it impossible to accurately identify who the voters. A better solution would have been to allow for the full number and enforce confidentiality rules on sensitive fields in voter records.

This four-digit constraint makes the social security number useful only to partially confirm the identity of a voter even when combined with other identifying information. States use the driver's license number as a key identifier, as it is administered by the state. Unfortunately not everyone has a driver's license, driver's licenses are not unique (one person many have many licenses), and driver's licenses cannot be used to follow a person when they move to a different state.

A recurring concern against the use of unique universal identifiers is that such a system would give the state government too much power. For example, there is concern that a state might bar citizens from voting for unrelated concerns, such as unpaid parking tickets. Such concerns are best addressed by legislation controlling the automatic exchange of information between different branches of government rather than limiting the technology used to effectively manage elections.

HAVA draws an important distinction between list maintenance that pertains to change of address and list maintenance that results in voter removal. Change of address operations are performed when voters inform the administration of the change (either directly or indirectly such as when updating their driver's license information). In that case the use of purge to remove voters who used to live in a state, but moved and registered to vote in another state is sensible. Another justifiable purge is that of prisoners incarcerated in the state where state law removes the right to vote from incarcerated felon. One important caveat is to allow for errors or mistakes on the part of the government to be corrected transparently and in timely manner as to prevent disenfranchisement, with minimal effort required on the part of the voter. One such system implemented in the Los Angeles County involves having voters certify and sign that they are the ones indicated on the record.

Federal laws recognize the risk for errors by requiring some sort of verification before removal from the rolls, but still leave many details to the discretion of state lawmakers. The resulting state and federal laws require much of elected officials, but often leave the details of their implementations to the Departments of State and in many cases to the Local Election Officials. For example, in Florida, although purges are required to be completed more than 90 days prior to any federal election, a voter may be removed *anytime* from a registration list if the voter has been convicted of a felony and has not had his voting right restored. As in other aspects of elections, the devil is in the details. Implementing sensible purges is quite a difficult endeavor, laid with potential for disenfranchisement.

During the 2000 presidential election, Florida came under the spotlight of the media for problems related to voting machines. On the other hand, the U.S. media largely overlooked the failure of Florida voter registration system. [Palast] detailed the decision by the Florida Department of State to implement a felon purge that resulted in disenfranchising 57,000 registered voters in a hotly contested election. The winning margin in Florida for the certified results was 537 votes. The Florida Department of State argued it designed the purge with broad matching criteria to capture as many as possible of the registered voters that were felons and thus barred from voting. The central problem with that decision is that it ignored the negative impact on thousands of voters who were lawfully registered to vote and wrongly removed them from the voter rolls without any due process. A disproportionate effect on Black voters occurred because Blacks are over-represented in the prison system, causing the purge to disadvantage political parties and candidates Black voters would be more likely to vote for.

For the 2004 presidential election, a lawsuit initiated by CNN and other news organizations and citizen associations forced the Florida Department of State to open its list of “potential felons” to public scrutiny well ahead of the general election. Here *potential felons* refer to voters on the voter registration database whose name somehow matched someone convicted of felony in the United States. Its plan to purge the statewide Florida voter registration database met fierce resistance when it became clear that Hispanic felons were practically excluded from the purge. The voter database kept track of the race and had a specific value for Hispanic. The felon list had also a race field, but did not have a specific value for Hispanic. Instead, Hispanics were classified as "White" in that list. The purge used the race field and required an exact match on the race to purge the record, thus excluding all Hispanics from the purge. It is important to note that conversely, excluding the race from the group of fields used for matching felons would have increased the false positive rate for all voters. On the other hand, requiring the race to match, but considering Hispanic in voter registration and White in felons list a match would increase false positive rate for Hispanic voters.

This problem illustrates well the risks and difficulties of matching records based on non-unique fields like the name of a person and date of birth as opposed to a unique identifier such as the full social security number. Without such unique identifier information, finding a matching strategy that does not disenfranchise a particular segment of the population is quite difficult. One important difference between removal of dead persons and removal of felons from voter rolls is that a much larger proportion of Blacks and Hispanics are convicted of felony than Caucasians. Death hits voters more blindly than justice in the United States. This is the main reason why the removal of felons from rolls is a controversial and politically charged question. Another important difference is that it is much easier to prove that one is alive than to prove that one is not a felon. A voter can easily prove her identity at the polling location, and thus be entitled to vote in this location even though the record indicates she is dead.

ChoicePoint, the company hired by the Florida Department of State to match names of a national felons list to the Florida VR in 2000, used a set of rules that included the first four letters of the first name, 80% match on the last name, and approximate date of birth [Palast]. These loose matching criteria, combined with the absence of public quantitative analysis for evaluating the risks of the approach, were central in creating the large-scale errors that occured. Time, budgetary, and political constraints may have contributed to the decision by Florida government official to use the resulting list of *potential felons* to remove voters from the rolls without additional case-by-case checking of the names and persons.

Experimental Study

In order to compare approximate and exact matching, the following hypothesis is considered: flexible matching based on approximate last name with the Soundex algorithm, exact first name, and exact date of birth does not result in accurate matches and creates large false positive errors (finding a match with the wrong person) when compared with exact matching on these three fields. This hypothesis was developed to illustrate the needs of a careful quantitative analysis for any sort of database maintenance.

Data Sources

The source for voter registration records was the Florida Voter Registration database (FLVR). It was obtained from the Florida Department of State, Division of Elections, and included updates to the database as of Aug. 15, 2004. The source for deaths records is from the website Rootsweb.com. This private website offers open and free access through the Internet to the Deaths Master File (DMF) from the Social Security Administration. The official distribution of the Death master file requires a paid subscription. Although Rootsweb.com is not endorsed by the SSA, its database seems accurate and relatively up-to-date. The website offers matching using the Soundex algorithm on the last name, as it helps uncover changes in names that can result from changed spelling during transcription of records on paper media, or from paper to digital media.

Filter programs

Filters are database programs that select records based on specific criteria. The filters used in this study select records by matching fields between records in FLVR with records in DMF following precise rules. Different filters have different rules for matching, as described below. The data processing consists of two steps. In the first step, the filter scans the FLVR database and for each record attempts to match that record to a record present in the DMF database. Matching records are written to a new file named after the source of data (the county) and an extension indicating which filter was used. In the second step, this file is then processed to extract each social security number from the file, and allow for faster comparison.
Other than the matching rules, the data processing is exactly the same.

Filter A: This filter requires an exact match of last name, first name, and date of birth to conclude to a match between a record in FLVR and a record in DMF.

Filter B: This filter requires an approximate match of the last name using the soundex algorithm, and requires an exact match for both first name and date of birth to conclude to a match between a record in FLVR and a record in SSD.

The only difference between the two filters is the use of the Soundex algorithm to match the last name in filter B versus an exact match for last name in filter A. The Soundex algorithm is a widely used method to categorize names by mean of a hash function. It allows for names with similar sounding to be classified together. The hash function reduces the name to a code made of the first letter of the name and a 3-digit number based on an English pronunciation of the name.

Both filters output a file of all matching records with information from both FLVR and SSD databases. The output file is then cleaned-up using a simple Perl script to a file with one record per line that includes the Social Security Number (SSN). Using the Unix command *wc*, the total frequencies for each filter are extracted from the file, and with the command *diff* the join frequencies are uncovered.

Results and Discussion

The experiments focused on three small counties of Florida. Table 1 shows the join frequencies found by scanning all voter records for Glades, Union, and Desoto Counties for a match with the SSD database using filters A and B.

	Glades County		**Union County**		**Desoto County**	
	Matches	%	Matches	%	Matches	%
Total	7301	100.0	7949	100.0	15191	100.0
A	234	3.2	121	1.5	166	1.1
B	270	3.6	150	1.9	240	1.6

Table 1: Matches for Filters A and B

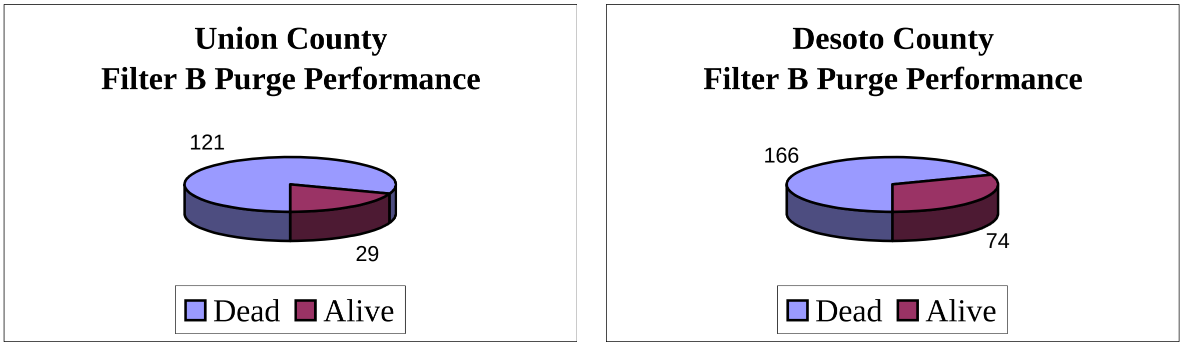

Figure 2a, b: Purge performed using approximate matching of names results in large errors.

It is assumed that filter A identifies perfectly dead persons, with no false positive or false negative errors. Of course, that may not be completely accurate, persons that have died may not have their death records entered yet in the database – these would be false negative for Filter A. Moreover, on a large base population (e.g., the whole state of Florida) it is impractical to obtain current and accurate records for all death certificates over the whole population. False positive for filter A are also likely as SSD cover the whole US population it is possible that two persons have the same exact first name, last name, and date of birth and that only one of them is dead. This simplifying assumption on filter A is still useful to show the risks of using broad matching rules.

These results indicate that the cost for using the broad filter B rather than the narrower filter A could disenfranchise 0.4 – 0.5% of the registered voters for the gain of removing 1-3% of records of actual dead persons. As shown on Figure 2a and 2b, the odds that the person identified by a record is alive, given that the person was flagged as dead by filter B ranges from one in eight to one in three, giving a very poor reliability to filter B.

In real applications, the filter A is not accessible to know exactly what are the true positive and true negative, and most likely some false negative would also exist. The filter B is broader than filter A and covers all of filter A; the positive count for filter B is larger than the count for filter A; the difference is made of an increase in the false positive and a decrease in the false negative. The increase in false positive is harmful and inevitable with a broad filter.

The decrease in false negative is what is usually used to justify the broader filter. Is this decrease significant? In these experiments on removal of dead voters, false negative can be attributed to two factors: delay of the death database updates and errors of spelling in the last name. Both filters are equally affected by the first factor as they use the same source for death certificates. Filter B could in theory catch some names of dead voters whose last name was misspelled either on the death certificate or in FLVR.

In order to estimate with confidence the false negative rate, a verification poll is necessary on a large sample of the population studied. This experiment covered over 30,000 registered voters; seventeen per thousand of the voters are dead according to filter A. Using a list generated with filter B or an even broader filter as a starting point, one could discover some of the false negative of filter A. However, only a comprehensive investigation would uncover all false negative cases. As the number of dead persons is already small even for filter A, the benefits hardly justify the extensive work.

The poor reliability of filter B is a direct result of the system used: matching a local list against a national list with no unique identifier. As one list covers a population an order of magnitude larger than another, the potential for false positive is very significant for any approximate, and even exact match on a limited number of field. The decision by the US Congress to limit SSN information to the last four digit of that number means that it must be used in conjunction with multiple other fields on exact matches to yield low false positive errors.

Conclusions

Removal of voters from voter registration lists, whether based on felons lists or death records, is an operation that has historically disenfranchised minorities and hurt the confidence in the election system. Any modification to the voter registration list should be balanced with both the gain and the problems caused by false positive. The argument that one wants to minimize false negative should never be used without sound experimentation. These procedures and algorithms should be open to public scrutiny and decided with a public debate. For each algorithm, a scientific study, including scientific sampling of records and manual verification of the decision should estimate the false positive rate that the purge may cause. Algorithms relying on broad match most surely increase the false positive rate but have no record of reducing false negative significantly.

One very important feature to include in the use of voter registration records is to allow a simple way for voters and the county to verify the validity of the database. One solution adopted by the Los Angeles County is to print out all voters in the voter registration database, active and inactive, in the same list in alphabetical order by precinct. If a voter who was flagged inactive shows up to vote, his signature is used as a statement certifying he is that voter. Election workers can also request additional information or documentation in accordance with the law as appropriate for this particular voter. By handling such error in the same manner as normal voters, the use of provisional ballots for errors originating in the list maintenance is greatly reduced.

Using modern database systems to maintain the voter registration list results in a very flexible system. One important feature is to keep inactive voters in the database. This way, inaccurate removals from the active list are more easily corrected. The cost of maintaining a database double the size of the number of registered and active voters is only a fraction more expensive. The incremental cost of maintaining a voter registration list larger than absolutely necessary is a very small price when compared to the social cost of voters who loose confidence in the election system.

Recommendations

- Publicly discuss algorithms and systems used for voter registration list maintenance, opening them to public scrutiny and
- Investigate the result of tentative voter purges using representative samples and careful verification of the reasons for voter on a significant and representative sample of the removal. This should be the primary
- Handle voter removal by changing their status in the voter registration database, but keeping them on the rolls available at polling place to make it an easy procedure to correct errors when the voter visits the polling place. Presuming that voters are innocent and allowing them to exercise their right until proven guilty is simply consistent with the general principle of justice in the United States.
- Require an independent confirmation of the reason for the definitive removal of records from the voter registration database, if such definitive removals are deemed necessary.
- Investigate occurrences of possible frauds and keep records that allow such investigation

and quantitative evaluation of fraud rather than speculations.

References

[NVRA 93] National Voter Registration Act of 1993, a.k.a. Motor Voter Act.
[HAVA 02] Help America Vote Act of 2002.
[Palast] Greg Palast, *The wrong way to fix elections*, Washington Post, 8 July 2001.