

Voter Removal from Registration List Based on Name Matching is Unreliable

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

October 28, 2004

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 everything from the budget to their staffing, office floor, and printing materials needs to conduct elections... Thus there is a strong public interest in keeping the VR up-to-date. This process is usually called voter registration list maintenance. This maintenance includes 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.

Administration of election is not uniform in the U.S. Each county administers election with much independence on the implementation details. Although centralization efforts 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 social security number. The importance of this information for the performance and accuracy of list maintenance operations is reviewed here. The law draws an important distinction between list maintenance that pertains to change of address and list maintenance that remove voters from the rolls.

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 must be completed more than 90 days prior to any federal election, a voter may be removed *anytime* from a registration list if the voter is 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

racial or other type of disenfranchisement. The risk to cause again large disenfranchisement over the 50 states as in the 2000 presidential elections is still very present. We first cover what went wrong in Florida as far as voter purges are concerned. We then detail a proposal for making purges more transparent and less threatening to the democracy.

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 black voters that were lawfully registered to vote and wrongly removed them from the voter rolls. The disproportionate effect on black voters occurred because Blacks are over-represented in the prison system.

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. 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. This problem illustrates well the risks and difficulties of matching records based on non-unique fields like the name of a person and his date of birth as opposed to unique identifier such as the social security number or the driving license number.

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.

As this particular example illustrates, 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, for socio-economic and systemic reasons, a much larger proportion of Black and Hispanics are convicted of felony. 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.

Experimental Study

In order to compare approximate and exact matching, the following hypothesis is considered: matching based on approximate (Soundex algorithm) last name, exact first name, and exact date of birth does not result in accurate matches and create potentially large false positive errors when compared with an algorithm that matches exactly on these three fields. The basis for the matching algorithms is realistic. 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% of the last name, and approximate date of birth [Palast].

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 Rootsweb.com. It offers access through the Internet to the Deaths master file from the Social Security Administration (SSA). 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 are the result of changed spelling.

Filter programs

Filters are used to match fields between records in FLVR with records in SSD. The filters may have different rules for matching, as described below. Other than the matching rules, the processing is exactly the same for all filters. The filter creates a new file named after the source of data (the county) and an extension indicating which filter was used. This file is then processed to extract each social security number from the file, and allow for faster comparison.

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 SSD.

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.

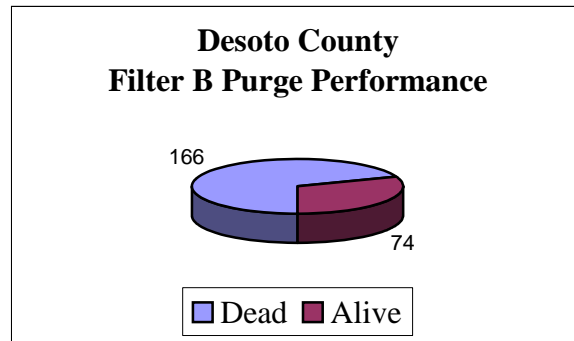
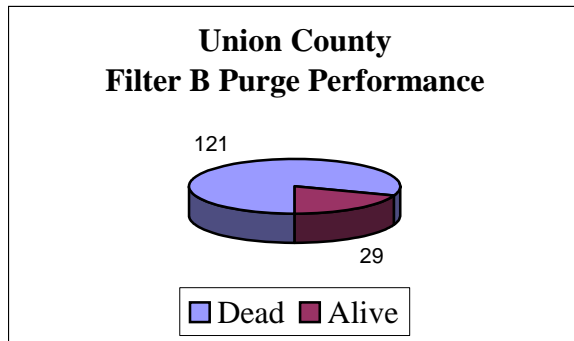
Results and Discussion

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		
	A	~A	Total	A	~A	Total	A	~A	Total
B	234	36	270	121	29	150	166	74	240
~B	0	7031	7031	0	7799	7799	0	14951	14951
Total	234	7067	7301	121	7828	7949	166	15025	15191

Table 1: Join Frequencies for Filters A and B

In Table 1, A indicates a record match found by filter A, and ~A indicates that a record is not matched by filter A. For example, in Glades County, 234 records match according to both filter A and B, but 36 records do not match according to filter A and match according to filter B. By design filter B matches every record that matches the SSD according to filter A.



First assume that filter A identifies perfectly dead persons, with no false positive or false negative errors. Of course, that may not be the case. E.g. 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.

This preliminary results indicate that the performance 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. In order to understand the trade-off involved,

an interesting measure is the probability that the person identified by a record is alive, given that the person was flagged as dead by filter B. This measure ranges from 13% to 31%.

Now relax the assumption that filter A matches perfectly dead persons. As explained above, 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. It is discussed with the previous assumption of a perfect filter A.

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. However, in the three counties studied here, none of the names had realistic spelling mistakes, only similar pronunciation.

In order to estimate with confidence the false negative rate, extensive field verification is necessary on a large sample of the population studied. This experiment covered over 30,000 registered voters; 17 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 this number is already small even for filter A, the benefits hardly justify the extensive work.

The high false positive rate is a direct result of the system used: matching a local list against a national list. As one list covers a population orders 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 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 small price when compared to the social cost of disenfranchising voters.

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 used at polling place to make it an easy procedure to correct errors when the voter visits the polling place.
- Require an independent confirmation of the reason for the definitive removal of records from the voter registration database, if such definitive removal are deemed necessary.
- Investigate occurrences of possible frauds and keep records that allow such investigation.

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.

Appendix

Background of statistical vocabulary

Given a population of size n with a proportion of that population of size m verifying a hypothesis

H , the ratio m/n is called the base rate for the hypothesis H .

A statistical test attempts to predict the membership of a particular subject. A positive test indicates that the subject verifies H . A negative test indicate the opposite. Depending on the results of the test for each subject of the population, we can draw the following cases:

- True positive: subject tests positive and verifies H .
- True negative: subject does not verify H and tests negative.
- False positive: the subject does not verify H but test positive. A.k.a. type I error.
- False negative: the subject verifies H but test negative. A.k.a. type II error.