Pennsylvania voter roll ID numbers

Preliminary Report

Andrew Paquette

10/13/2024

Introduction

The investigation of Pennsylvania voter rolls was prompted by findings in other states:

New York: An estimated 2 million illegal "clone" records were discovered, along with four algorithms used in ID assignment. These algorithms can predict voter status, identify clones, reveal deleted SIDs, and add hidden attributes to records (Paquette 2023).

One algorithm predicts voter status with 99.34% accuracy and can also identify clone records. Another algorithm reveals deleted SID numbers and who those numbers were originally assigned to. This second algorithm, called “The Spiral,” also adds an attribute to all records that is effectively a third, and very well-hidden, ID number. An Algorithm ID (AID).

New Jersey: An encoded identification system that transforms and reverses ID numbers was found, potentially allowing covert record identification (Paquette, in press).

Ohio and Texas: Hidden attributes in voter records enable covert tracking in populous counties.

Hawaii: A tagging mechanism on UUID numbers segregated 10% of records, which have since been deleted.

These findings suggest the possibility of hidden attributes in voter roll data fields, particularly in unique identifiers like State ID (SID) and County ID (CID) numbers. This led to the current investigation of Pennsylvania's voter rolls.

A fundamental rule of database management is that all data should be transparent, traceable, and used only for its intended purpose. The algorithms found in various state databases violate this rule by introducing what amounts to undocumented attributes into the database. This makes it untraceable by normal means and can enable manipulations that violate the intended purpose of the databases.

Widespread credible reports of election fraud in Pennsylvania led me to ask for their voter rolls in 2022. In my initial analysis, I did not find any sign of unusual algorithms, though there were tens of thousands of suspicious records. More recently, I obtained another copy of the Pennsylvania rolls, dated 9/8/2024, for a more detailed analysis.

This preliminary report aims to:

1. Determine if hidden attributes exist in Pennsylvania's voter roll records.
2. Assess whether these attributes were generated by deterministic (predictable) algorithms.

The focus is not on whether "an algorithm was used" - all ID assignment software uses algorithms. Instead, we're interested in algorithms that:

- Are over-engineered for ID assignment
- Employ obfuscation techniques
- Add hidden attributes or segregate data

While time constraints prevent a full solution of any algorithms found (unlike in NY), their presence and purpose can be demonstrated without complete reversal.

Initial results reveal 115,434 cloned records in Pennsylvania's current database, a number sufficient to justify ID-tracking algorithms. Nine of Pennsylvania's 67 counties employ a complex algorithm mapping Legacy ID (LID) numbers to ID numbers. Notably, this includes Allegheny and Philadelphia counties, which together account for 22.45% of all registrations.

Preliminary results

The quickest way to establish whether records appear normal is to compare ID1 (state ID) with either ID2 (county ID) or Registration Date (RegDate), determining if both ascend at similar rates. This analysis uses three key steps:

1. Scatterplots: Used to quickly segregate "normal" counties from those with anomalous patterns.
2. Gap value analysis: Examines the differences between adjacent ID numbers.
3. Gap frequency analysis: Identifies any recurring mathematical patterns in the ID numbers.

Utilizing these methods, nine counties were identified for further investigation: Allegheny, Somerset, Fayette, Schuylkill, Northampton, Columbia, Lycoming, Lawrence, and Philadelphia.

Scatterplots

For the first stage of this analysis, Pennsylvania’s voter ID numbers were compared to registration dates. In a scatter plot, sequential assignment creates a sloped line starting at the lower left corner of the graph (the lowest ID numbers) and ascend toward the upper right (highest numbers).

In 61 of Pennsylvania’s 67 counties, scatterplots show that increasing ID numbers correlate with more recent registration dates. While there is a large gap in the numbers, the crucial observation is that numbers ascend as expected on both sides of this gap (Figure 1). These patterns are considered "normal" for this stage of the analysis. However, they do deviate from standard database management due to sometimes large gaps between sequential numbers.

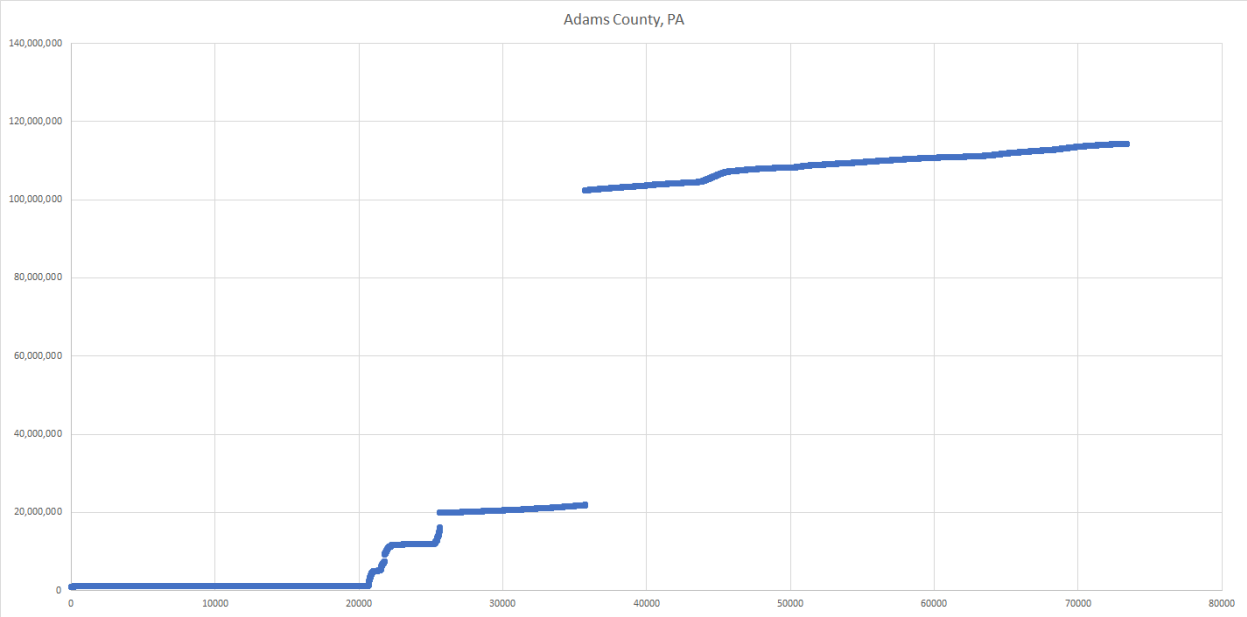


Figure 1 Adams County, PA, scatterplot (X: RegDate, Y: ID Num)

Five PA counties (Fulton, Greene, Juniata, Lycoming, and Philadelphia) exhibit unusual clustering in ID numbers for registrations before 2009. In these counties, pre-2009 ID numbers don't correlate well with registration dates, instead using the same ID ranges across multiple years.

Philadelphia County's ID numbers are divided into seven distinct clusters, illustrating the complexity of the assignment system: (Figure 2).

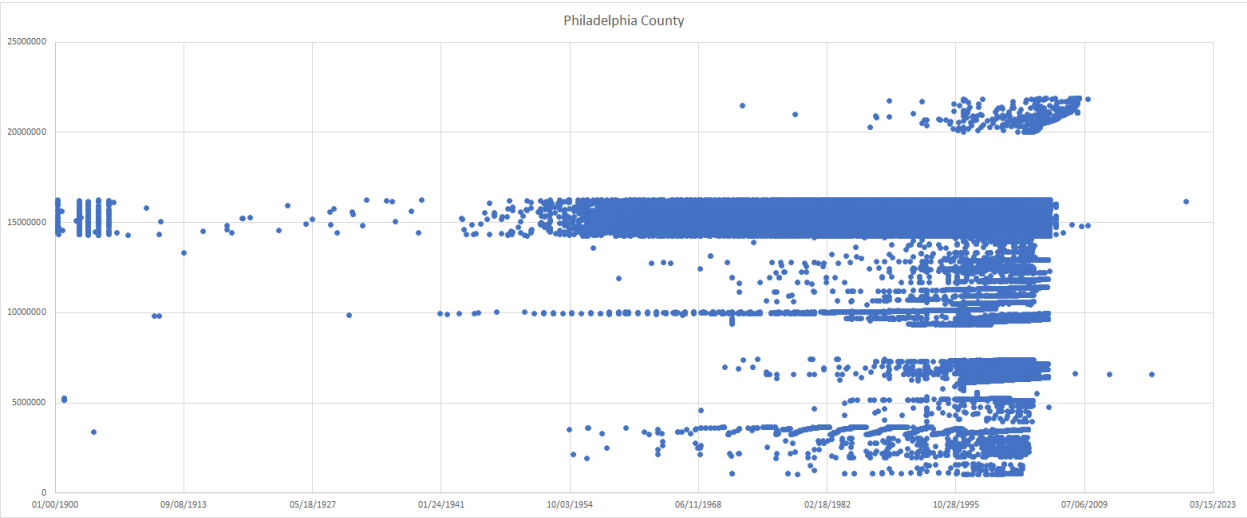


Figure 2 Philadelphia County, close-up

Table 1 Breakdown, Philadelphia ID numbers

Cluster	Date Range	ID Range	Records	Key Features
1	Pre-1920	14.25M-16.25M	236	Full ID range used
2	1920-2004	1.02M-4.05M	4,611	Horizontal and vertical stripe patterns
3	1968-2005	4.10M-6.98M	12,416	Horizontal stripes, complex formation around 1995
4	1931-2005	7.01M-13.99M	12,414	27 overlapping horizontal stripes
5	1920-2005	14.02M-16.00M	316,792	Vertical stripes with strict date segregation
6	1986-2008	20.00M -	106,645	Two horizontal stripes, transitioning to post-2008 pattern
7	Post-2008	N/A	567,002	Ascending IDs correlating to registration dates, with large gaps

This complex, multi-layered system deviates significantly from standard sequential ID assignment practices, suggesting a sophisticated and potentially unnecessary level of complexity in voter record management. To implement such a system, programmers would need to:

1. Manage multiple ID assignment strategies simultaneously.
2. Develop logic to choose between strategies for each ID assignment.
3. Handle overlaps in both time ranges and ID ranges across clusters.
4. Maintain records to prevent ID duplication within this complex structure.
5. Implement triggers for abrupt changes in assignment strategies.
6. Create algorithms for pseudo-random ID distribution in certain ranges.

This approach introduces several challenges:

1. Increased likelihood of implementation errors due to complex logic.
2. Potential performance issues from maintaining records across overlapping clusters.
3. Difficulties in maintenance and future modifications.
4. Less straightforward correlation between IDs and registration dates.
5. Higher computational resource requirements.
6. Additional complications for data integrity verification and auditing.

The transition to a more conventional system post-2008 aligns the database more closely with common practices.

Several explanations for this structure are possible:

1. ID ranges assigned in different eras, combined with voter movement within the state.
2. Segregation of records based on hidden attributes, which would be considered unethical.
3. Reservation of unused space for future registrations, though this is unnecessary with standard sequential numbering.
4. System evolution, with database designers changing ID assignment methods multiple times.

Standard sequential numbering would have offered numerous advantages: ease of implementation, transparency, usability, cost-effectiveness, and reduced administrative overhead. The complexity of the observed system raises questions about its purpose and necessity, particularly given the large number space available in Pennsylvania counties

Gap analysis

Gap analysis compares two values by subtracting one from the other to determine the difference. In New York, this method revealed a 'Spiral' algorithm using Repunit-based patterns (e.g., 1,111, 111, 11) in SID numbers. This pattern was consistent, with predictable variations for missing records. Gap analysis can be affected by added or deleted records, changing gap values unpredictably.

New York's gap distribution was highly organized, following a predictable pattern: every 10th gap was 11, every 100th was 111, every 1000th was 1,111, and so on. Other predictable values were interspersed.

In Pennsylvania, ID numbers for all 67 counties were analyzed, with more extensive checks for larger counties. While significant gaps exist between adjacent numbers when sorted ascending, no definable patterns emerged. However, certain gap values recurred more frequently than expected in a random assignment, suggesting a non-random process. A gap frequency analysis was conducted to further investigate these high-recurrence values.

Gap sizes in a natural distribution are inherently biased by the mathematics of subtraction. Smaller gaps occur more frequently because they can be produced by a wider range of number combinations. Conversely, larger gaps are rarer, as fewer number pairs can produce them when subtracted. This relationship between gap size and frequency is illustrated in *Table 2*, showing an inverse correlation between gap magnitude and its likelihood of occurrence.

Table 2 Subtraction product distribution based on number pair combinations

0	1	2	3	4	5	6	7	Gap	Total combos	Pct
1	0	NA	NA	NA	NA	NA	NA	0	7	25.00%
2	1	0	NA	NA	NA	NA	NA	1	6	21.43%
3	2	1	0	NA	NA	NA	NA	2	5	17.86%
4	3	2	1	0	NA	NA	NA	3	4	14.29%
5	4	3	2	1	0	NA	NA	4	3	10.71%
6	5	4	3	2	1	0	NA	5	2	7.14%
7	6	5	4	3	2	1	0	6	1	3.57%

The general expectation is that in a random distribution of numbers, gap frequencies would follow this natural pattern. Most Ohio counties demonstrate this expected descending gap frequency distribution (Table 3), with only three exceptions (two illustrated-Lucas and Montgomery Counties).

Table 3 Sample Ohio counties show descending gap frequencies, with 2 exceptions in Lucas and Montgomery Counties

CID Gap Value	Lawrence Frequency	Licking Frequency	Logan Frequency	Lorain Frequency	Lucas Frequency	Madison Frequency	Mahoning Frequency	Marion Frequency	Medina Frequency	Meigs Frequency	Mercer Frequency	Miami Frequency	Monroe Frequency	Montgomery Frequency
1	24,297	68,766	17,901	126,786	121,830	13,898	81,955	21,284	74,315	8,224	17,125	40,114	5,105	149,615
2	7,174	22,141	6,210	39,467	35,113	4,362	28,763	8,167	22,661	2,846	5,489	13,296	1,794	47,989
3	3,856	11,651	3,246	19,672	15,591	2,370	15,548	4,135	11,728	1,570	2,811	7,223	1,002	23,378
4	2,370	6,899	1,805	11,137	7,691	1,480	9,711	2,450	7,005	930	1,539	4,227	592	12,150
5	1,616	4,471	1,096	6,699	4,261	1,015	6,486	1,528	4,382	599	915	2,908	341	6,870
6	1,146	2,787	684	4,209	2,388	705	4,447	912	2,897	370	559	1,951	248	4,015
7	815	1,921	447	2,761	1,476	500	3,341	562	1,962	247	330	1,438	180	2,511
8	629	1,374	257	1,924	28,301	366	2,531	417	1,378	176	220	953	83	20,905
9	464	961	184	1,360	526	274	1,891	260	954	117	125	704	82	1,403
10	377	707	100	939	320	199	1,570	158	706	97	77	527	41	601
11	282	483	59	696	299	142	1,312	144	517	54	54	395	29	660
12	217	368	42	539	144	122	1,071	78	384	53	32	315	20	366
13	187	258	35	411	94	86	859	41	273	31	16	277	12	263

Gap value frequencies were calculated for gap values 1 through 100 in every Pennsylvania county. The results revealed a remarkably consistent pattern across all counties, markedly different from the findings in Ohio. In Pennsylvania, the gap value of 1 consistently has the highest frequency, with each subsequent value decreasing smoothly and predictably.

This contrasts sharply with Ohio, where some counties (like Lucas and Montgomery) showed dramatic spikes in certain gap values. For instance, in Ohio's Lucas County, the frequency jumps from 1,476 for gap 7 to 28,301 for gap 8 - a nearly 20-fold increase. Pennsylvania's data, however, shows no such anomalies. Even in larger counties like Allegheny, the decrease in frequency from one gap value to the next is gradual and consistent.

This uniform pattern across Pennsylvania counties suggests a standardized approach to ID assignment, unlike the varied and sometimes anomalous methods apparent in Ohio. The absence of significant spikes or irregularities in Pennsylvania's gap value frequencies indicates a more systematic and consistent ID assignment process throughout the state.

Number Analysis

The analysis of Pennsylvania's voter registration data revealed two contrasting patterns. Most counties exhibited a nearly uniform distribution of ID numbers, unlike the anomalies found in Ohio. However, a few counties showed overlapping date and ID number ranges, reminiscent of the complex patterns observed in New York.

This contrast suggested that potential algorithms might be obscured within the broader data set. A closer examination of Pennsylvania's 8,846,634 registrations uncovered a significant pattern: all 67 counties with Legacy ID (LID) numbers were assigned specific ID number ranges (Table 4). Conversely, counties without LIDs showed no clear county-specific ID number allocation.

Table 4 Example, county ranges, Pennsylvania

County ID	MIN ID	MAX ID	Gap to Prev County	Range	Record Count All	Record Count with LegID	Pct Used	Leg ID Type
21	1,000,001	1,129,783		129,783	64,940	54,867	50.04%	Alpha
1	1,215,002	1,266,200	85,219	51,199	24,184	20,030	47.24%	Decimal
4	1,300,029	1,414,319	33,829	114,291	53,311	46,676	46.64%	Decimal
10	1,496,001	1,613,771	81,682	117,771	54,441	45,825	46.23%	Alpha/Decimal
2	1,613,772	3,077,983	1	1,464,212	423,639	365,715	28.93%	X10
9	3,254,002	3,670,093	176,019	416,092	208,520	179,025	50.11%	Alpha
3	3,944,005	3,986,993	273,912	42,989	21,842	17,985	50.81%	Decimal
63	4,018,001	4,162,449	31,008	144,449	62,080	52,239	42.98%	Decimal
30	4,257,004	4,282,374	94,555	25,371	11,111	9,529	43.79%	Decimal
56	4,298,004	4,348,023	15,630	50,020	25,256	21,929	50.49%	Decimal
26	4,383,003	4,489,101	34,980	106,099	41,871	36,083	39.46%	Decimal

This finding indicates that LID presence is a key factor in ID number assignment across Pennsylvania counties, potentially concealing more subtle algorithmic patterns. The situation parallels the New York case, where data filtering was necessary to reveal underlying algorithms.

After filtering for records with Legacy ID (LID) numbers, a new gap and frequency analysis revealed a mapping algorithm in nine Pennsylvania counties: Allegheny, Somerset, Fayette, Shuylkill, Northampton, Columbia, Lycoming, Lawrence, and Philadelphia.

In these counties, LID numbers are grouped by their last digit when sorted by ID (Table 5). For example, in Philadelphia County:

1. LID numbers ending in "0" (like 1,010,005,440) form a contiguous block of 35,699 records.
2. This is followed by a block of 35,560 LIDs ending in "1" (starting with 1,012,507,251).
3. The pattern continues through digits 2-9.
4. In total, 359,212 LID numbers are assigned to sequential ID numbers in this manner.

This systematic grouping suggests a deliberate algorithmic approach to mapping LIDs to IDs, distinct from the more random distribution observed in other counties.

Table 5 Philadelphia County LID numbers grouped by last digit

ID Num	Legacy ID (0)	ID Num	Legacy ID (1)	ID Num	Legacy ID (2)	ID Num	Legacy ID (3)	ID Num	Legacy ID (4)
14,245,005	1,010,005,440	14,446,159	1,012,507,251	14,647,217	1,012,194,922	14,848,303	1,011,070,133	15,050,065	1,011,070,134
14,245,008	1,012,065,580	14,446,160	1,012,511,471	14,647,221	1,012,577,872	14,848,306	1,012,179,583	15,050,076	1,012,548,984
14,245,032	1,013,330,930	14,446,169	1,013,055,171	14,647,227	1,013,019,682	14,848,308	1,012,577,873	15,050,081	1,012,881,474
14,245,036	1,013,512,060	14,446,172	1,013,170,471	14,647,230	1,013,170,432	14,848,309	1,012,580,573	15,050,084	1,013,100,114
14,245,042	1,013,693,930	14,446,179	1,013,680,001	14,647,255	1,013,926,872	14,848,311	1,012,855,413	15,050,090	1,013,306,714
14,245,045	1,013,734,490	14,446,180	1,013,727,901	14,647,265	1,014,120,652	14,848,315	1,013,078,943	15,050,101	1,013,550,224
14,245,066	1,014,175,510	14,446,182	1,013,843,451	14,647,275	1,014,336,092	14,848,317	1,013,154,293	15,050,103	1,013,713,894
14,245,079	1,014,502,560	14,446,183	1,013,849,711	14,647,277	1,014,447,562	14,848,325	1,013,525,843	15,050,105	1,013,728,434
14,245,082	1,014,536,260	14,446,191	1,013,985,471	14,647,280	1,014,481,752	14,848,329	1,013,694,273	15,050,109	1,013,919,754
14,245,083	1,014,539,500	14,446,194	1,014,043,071	14,647,282	1,014,518,512	14,848,353	1,014,150,073	15,050,110	1,013,933,404

The distribution of LID last digits is remarkably uniform, with each digit occurring at a frequency within 1% of the others. This near-equal distribution masks the underlying pattern when only frequency data is examined. The contiguous blocking by last digit is not evident from frequency analysis alone.

The pattern only becomes apparent when:

1. Data is filtered to include only records with LIDs
2. Records are sorted by ID (not LID) in ascending order
3. A visual inspection of the sorted data is performed

A simple count or frequency analysis of these values would not reveal the underlying structured assignment. This structure, whether intentional or not, has significant implications for data analysis and transparency. It underscores the importance of thorough, multi-dimensional analysis in uncovering patterns in complex datasets (Table 6).

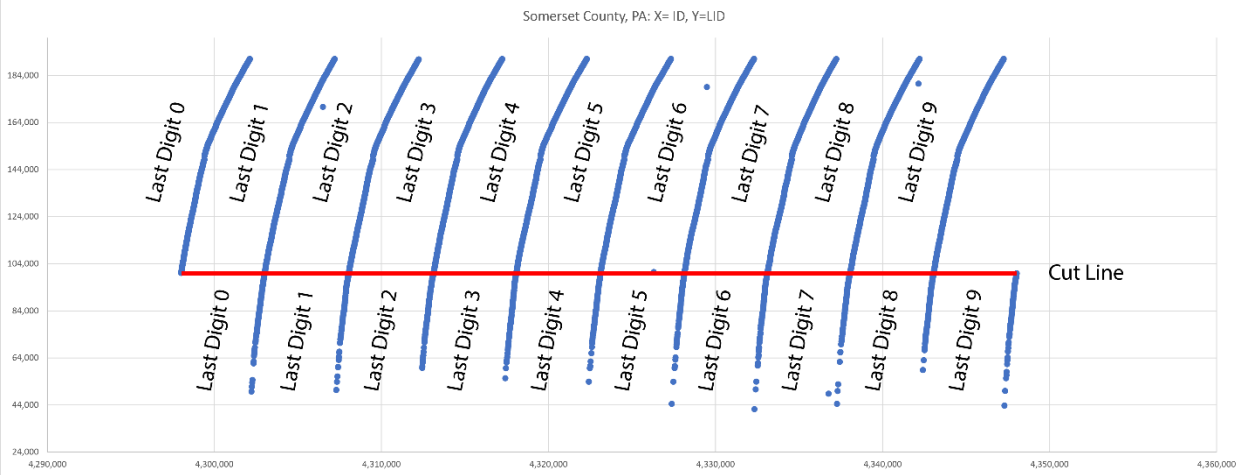
Table 6 Philadelphia County, PA, count of each last digit occurrence

LEG ID Last Digit	Freq Leg ID Last Digit
0	35,700
1	35,560
2	35,820
3	35,959
4	36,009
5	36,026
6	36,070
7	35,929
8	36,114
9	36,025
	359,212

Grouping LID numbers by last digits is similar to an artifact of New York's "Spiral" algorithm, which sort State ID (SID) numbers into power of ten groups. In New York, this has a strong obfuscatory effect due to subsequent transformations performed by the algorithm.

In Pennsylvania, Somerset County introduces another layer of complexity, by shifting the position of last digit groups at a consistent position (about ID 100,000). Because of this, each group of contiguous ID numbers contains two, not one set of matching last digits (Table 7). Although found in only one county, this is intriguing because of its similarity to New York's "Spiral" algorithm, which also has a cut and shift feature that offsets the start and end points of the numbers in each power of ten column.

Table 7 Somerset County cut and shift artifact



The relationship between ID numbers and Legacy ID (LID) numbers in some Pennsylvania counties reveals an interesting pattern: LID numbers are mapped to ID numbers in lexicographic order, rather than numerical order (Table 8). This means:

1. LID numbers are sorted as if they were text strings, not numerical values.
2. This creates an apparent "decimalized" sorting effect, though no actual decimal conversion occurs.

For example, in the image:

- LID 152 precedes 1,522
- LID 16 comes after 1,562 but before 161

This lexicographic sorting results in:

- Single-digit LIDs appearing later in the sequence (e.g., 16)
- Multi-digit LIDs being ordered based on their first digit, regardless of total value

Table 8 Adams County, PA, lexicographic sort of LID numbers resembles decimalized sort order

ID Num	Legacy ID	Decimalized	RegDate
1,250,737	148,189	0.148189	01/03/2003
1,250,740	148,192	0.148192	01/03/2003
1,250,761	**152**	**0.152000**	**09/17/1962**
1,250,766	1,522	0.152200	07/15/1975
1,250,771	**15,278**	**0.152780**	**09/30/1944**
1,250,774	154	0.154000	09/17/1962
1,250,777	156	0.156000	02/28/1963
1,250,778	1,561	0.156100	07/07/1976
1,250,779	1,562	0.156200	07/07/1976
1,250,807	**16**	**0.160000**	**06/08/1964**
1,250,812	161	0.161000	08/09/1962
1,250,822	165	0.165000	08/10/1962

While lexicographic sorting is a valid method, especially when dealing with mixed alphanumeric and numeric identifiers, its application here creates an unintuitive ordering from a numerical perspective. The fact that 34 counties sort their numbers numerically, while others use this lexicographic approach, introduces inconsistency across the state's database management practices.

This mixed approach raises questions about:

1. The rationale behind using different sorting methods across counties
2. Potential implications for data analysis and integration
3. Whether this inconsistency serves a specific purpose or is an artifact of varied database management practices

The complex patterns observed in Pennsylvania's voter ID system create a structure that could potentially allow for the assignment of hidden attributes to voter records. The segregation of LID numbers by last digit, overlapping date and ID number ranges, inconsistent use of lexicographic sorting, and county-specific anomalies like Somerset's shift pattern all contribute to a multi-layered system. This complexity enables the possibility of covert categorization without adding visible fields to the database.

The overall deviation from standard sequential ID assignment practices allows for more complex data structures, which could theoretically encode additional information not apparent in the visible fields. These non-standard practices create a system where hidden attributes could be assigned and tracked without obvious indicators in the database.

Suspicious records

Complex voter roll ID assignment systems are typically justified by a need that cannot be easily managed through simpler methods. In public databases, a straightforward approach often suffices: serial numbers incrementing by a fixed value as records are added. More intricate systems, if legitimate, should produce outputs that are transparent and easily understood.

The presence of a significant number of records requiring covert management could potentially justify a more complex system. In New York and Wisconsin, for instance, the high number of clone records (estimated 2,000,000 and 500,000 respectively) might warrant such complexity. However, Pennsylvania's situation differs markedly.

Pennsylvania's voter database contains 8,846,634 records, of which only 9,924-19,846 (about 0.11%) are potential clones. These were identified by matching last name, first name, and date of birth - a method aligned with many state election regulations for initial duplicate checks.

While this identification method may produce some false positives, the combination of full name and complete birth date sharing is statistically rare. The most common exception involves immigrants with missing birth date information who are assigned default dates.

The number of clones in Pennsylvania, while problematic, may not alone justify the implementation of a complex voter tracking algorithm. However, this doesn't preclude other reasons for such a system:

1. There may be additional suspicious records not identified by this analysis.
2. The system could be preemptively positioned for future use or expansion.
3. Other administrative or operational factors unknown to outside observers might necessitate this complexity.

Analysis of Pennsylvania's voter records revealed several anomalies:

- 26,453 registrations predate the voter's birth
- 29,707 voters with recorded votes before their 15th birthday
- 85 cases where the last vote date precedes the birth date

These anomalies are primarily concentrated in Blair, Dauphin, and Venango counties.

Of greater concern are 228,997 active records with no activity since 1/1/2018. Pennsylvania law requires inactive status for voters after missing two successive federal elections. Since 2018, three such elections have occurred (2018, 2020, 2022), meaning these records should be inactive or removed.

These outdated active records are mainly found in:

- Philadelphia: 39,856
- Allegheny: 24,065
- Montgomery: 12,619
- Bucks: 11,170
- Delaware: 9,217

Ignoring voter status, 514,422 records show no activity since 1/1/2018. While Pennsylvania lacks a strict removal timeline, federal law allows removal after two federal elections without activity, following an unanswered confirmation notice. The persistence of these records through three election cycles raises questions about list maintenance procedures.

These findings suggest potential systemic issues in voter roll management and warrant further investigation.

Comments

This study of Pennsylvania's voter rolls reveals evidence of multiple ID number assignment algorithms. These appear overly complex, potentially enabling data segregation and hidden attribute assignment. The presence of 514,422 apparently inactive records for 6 years exceeds normal error rates or acceptable administrative standards. Such a large number of problematic records could potentially impact election outcomes if manipulated.

These findings suggest potentially problematic management of Pennsylvania's voter roll records. The algorithm's use creates a hidden classification system for data segregation, posing a security risk. The high number of questionable records exacerbates this risk, as they could be targets for voter roll misuse - a concern recently realized when Wisconsin mailed absentee ballots to inactive voters.

Pennsylvania should investigate:

- When and by whom the algorithm was introduced
- Its intended purpose
- Associated costs
- Prior awareness among officials

Additionally, Pennsylvania should consider removing all inactive records and those incorrectly marked as active. Retaining unusable voting records serves no legitimate purpose. If preserving voter history is a concern, these records could be archived separately from the active rolls.

This investigation is crucial for maintaining electoral integrity, data security, and public trust in the democratic process.

References

Paquette, A. (2023). "The Caesar cipher and stacking the deck in New York State voter rolls " Journal of Information Warfare **22**(2): 86-105.

Paquette, A. (2024). "New Jersey voter ID numbers reconfigured with shift cipher." (In-Press).