Preliminary report: Analysis of Ohio voter roll ID numbers

Andrew Paquette

9/14/2024

Introduction

In the United States, voter roll databases are public under the law. For that reason, privacy and security concerns are different than in other industries, like banking or insurance.

The two most common data privacy measures are encryption (obfuscation) and masking. Encryption modifies data so that it is not recognizably related to the original input (Templ and Sariyar 2022). For instance, by changing the name “Smith” to “2%dfV+”. Masking is what credit card companies do when they conceal all but the last four digits of their customer’s account numbers.

In voter rolls, encryption and masking are not expected because both defeat the purpose of a publicly available database. In addition, those methods may violate legal requirements for public disclosure (1993). Voter registration laws at the federal and state level prohibit the creation of records that can be used to vote more than once. If followed, those laws would make it impossible to create most types of illegal records, and would dramatically reduce administrative or clerical error.

One such requirement in New York (NYBOE 2022) is that new registration applications must be checked against existing records to determine if the would-be registrant is already registered. To do this, a clerk must check for matching names. If found, they must search matching birthdates. If that is also found, they must check either driver’s license number or the last four digits of their social security number. If a match is found, the existing registration must be updated rather than creating a new one.

Unless the applicant is using fraudulent identification, these measures are sufficient to prevent accidental processing of multiple applications for the same individual. Another easily implemented error prevention technique is to use data validation. This is built into all commercially available databases, and is generally built into custom database solutions as well. Data validation prevents the entry of obviously incorrect data, such as a three-digit zip code, phone numbers containing letters, birth dates in the remote past or the future, etc.

Clearly, these tools are not universally in use. I began my study of voter rolls in New York, as research director of New York Citizen’s Audit (NYCA). Within months, we’d discovered approximately two million cloned records. These are two or more unique state ID numbers attached to the same individual, creating the possibility of multiple voting. Of some concern, about half a million such records were found due to well-hidden encryption-like algorithms buried in the method used to assign voter ID numbers (Paquette 2023).

One of those algorithms allowed me to find deleted clone records and to identify who those records originally belonged to (Table 1). That shouldn’t be possible. Another algorithm allowed me to predict voter status with approximately 99.34% accuracy based on ID number alone.

instead of looking up SBOEID 20,000,000, it uses the mapping table to find fraudulent record 27, which is the same as SBOEID 20,000,000. No one else would know that, because illegitimate record 20,000,000 looks the same as all the legitimate records.

This alternate means of identifying records of interest allows those records to reside in plain sight with significantly reduced risk of discovery. This is similar to steganography, a form of encryption that disguises the fact encryption was used (Kaur and Rani 2016).

New Jersey's voter rolls contain clear evidence of algorithmic manipulation. Their ID numbers have been treated as modular components that that can be reorganized to change their order. As in NY, knowledge of how this was done creates a way to covertly track records.

Researcher Vico Bertogli discovered that slightly less than 10% of Hawaii's voter ID numbers share the same final 12 digits in a 32-digit UUID. This is enough to uniquely identify those records as belonging to a group. Scatterplots of North Carolina and Ohio voter IDs show signs of inconsistent or unusual methods used to assign ID numbers.

It is not possible to know from the available data whether any of these algorithms, or Hawaii's tagging method, have been designed for, or used, nefariously. New York and New Jersey's algorithms have no readily apparent legitimate justification for being in a voter roll database. They don't improve efficiency, nor do they protect any data. Sensitive information like social security and driver's license numbers is simply withheld, eliminating the need for encryption or masking. Similarly, Hawaii's modified ID numbers have no apparent benefit.

My study of the New York and New Jersey algorithms is advanced enough that both could be described as complete. I have only begun to look at North Carolina and Ohio. However, I can say that both states have used at least two different algorithms in their rolls to assign voter ID numbers. Their presence alone is interesting, because custom-designed methods for assigning ID numbers are unnecessary in a public database.

All databases come with tools to automatically assign ID numbers in chronological order. Absent any compelling reason to do it differently, this option is generally the least expensive, easiest to use, and most transparent. This was not done in at least some Ohio counties, leading to the question, why?

Ohio overview

Ohio's counties range from showing no notable irregularities to exhibiting differences worthy of further investigation. The most intriguing counties thus far are Cuyahoga, Franklin, Lucas, and Montgomery.

The principal question asked for this analysis is, "Do Ohio's voter rolls exhibit evidence of algorithmic manipulation for covert tagging or selective data obfuscation?"

Secondarily, it seeks to learn whether there is a sufficient number of suspicious records to make the use of such an algorithm worthwhile.

This analysis serves as a starting point for further investigation. While not exhaustive, it highlights potential anomalies that warrant closer examination. Given the importance of election security, thorough scrutiny is essential. Additional findings are likely with continued research.

Table 1 Missing records in this table of NY registrations were likely deleted clone records belonging to identifiable voters

STATE ID Num	State ID Gap	State ID Distance to MIN	State ID Distance to MAX	Power	County ID NUM	CID Decimalized	DOB	Status	VR Source	County Code	Reg Date	Record ID
21,942,837	-50,381	83	50,659	100	1,000,082	0.10000820	5/29/1944	P	CBOE	10	6/1/2007	3,473,759
21,942,948	111	194	50,548	100	1,000,083	0.10000830	12/9/1983	P	CBOE	10	5/25/2007	3,473,870
MISSING				100	1,000,084							
21,943,170	#VALUE!	416	50,326	100	1,000,085	0.10000850	10/5/1988	A	CBOE	10	5/25/2007	3,474,092
21,943,280	110	526	50,216	100	1,000,086	0.10000860	8/22/1977	A	CBOE	10	5/30/2007	3,474,202
MISSING				100	1,000,087							
21,943,502	#VALUE!	748	49,994	100	1,000,088	0.10000880	12/29/1987	A	CBOE	10	5/25/2007	3,474,424
MISSING				100	1,000,089							
21,943,725	#VALUE!	971	49,771	100	100,009	0.10000900	4/6/1947	A	CBOE	10	1/21/1980	3,474,647
MISSING				100	1,000,090							
21,943,947	#VALUE!	1,193	49,549	100	1,000,091	0.10000910	5/20/1949	P	CBOE	10	6/1/2007	3,474,869
21,944,058	111	1,304	49,438	100	1,000,092	0.10000920	8/16/1966	P	CBOE	10	6/4/2007	3,474,980
21,944,169	111	1,415	49,327	100	1,000,093	0.10000930	4/29/1972	P	CBOE	10	6/4/2007	3,475,091
21,944,280	111	1,526	49,216	100	1,000,094	0.10000940	2/26/1985	P	CBOE	10	6/4/2007	3,475,202
21,944,390	110	1,636	49,106	100	1,000,095	0.10000950	8/28/1957	A	CBOE	10	6/4/2007	3,475,312
21,944,501	111	1,747	48,995	100	1,000,096	0.10000960	8/13/1986	P	CBOE	10	6/4/2007	3,475,423
MISSING				100	1,000,097							
21,944,724	#VALUE!	1,970	48,772	100	1,000,098	0.10000980	3/6/1986	P	CBOE	10	6/4/2007	3,475,646
21,944,835	111	2,081	48,661	100	10,001	0.10001000	10/13/1966	A	CBOE	10	9/24/1996	3,475,757
21,944,946	111	2,192	48,550	100	100,010	0.10001000	5/17/1950	P	CBOE	10	1/22/1980	3,475,868
21,945,057	111	2,303	48,439	100	1,000,101	0.10001010	9/27/1983	P	CBOE	10	6/4/2007	3,475,979
MISSING				100	1,000,102							

One of New York's voter roll algorithms does not encrypt or mask any information. Instead, it cleverly associates existing data with an extremely complex deterministic mapping scheme. This algorithm, which I call the "Spiral," decides which County ID (CID) numbers are attached to which State ID numbers (SBOEID). In so doing, it creates something called a "mapping table" that allows covert interaction with records by referencing the table instead of the public-facing ID numbers (Table 2). Board of Elections employees would have no access to the mapping table without knowledge of the algorithm.

Table 2 Spiral algorithm number scrambling method, simplified. "AID" is the algorithm-determined rank ID for each record.

PROJECTED	0	100,000	10,000	1,000	100	10	1
MIN	35,080,636	35,163,969	35,088,969	35,081,469	35,080,719	35,080,644	35,080,637
MAX	36,049,986	36,022,208	36,047,208	36,049,708	36,049,958	36,049,983	36,049,986
Range	969,351	969,350	969,342	969,255	968,383	959,659	872,418
Repunit	NA	111,111	11,111	1,111	111	11	1
Count	1	8	87	872	8,724	87,241	872,418
Adjusted range	969,350	969,342	969,255	968,383	959,659	872,418	0
AID First	1	2	84	834	8,334	83,334	833,334
AID High (CUT)	NA	NA	96	968	9,692	96,933	969,351
AID Low (CUT)			10	97	969	9,693	96,934
AID Last	1	9	83	833	8,333	83,333	833,333
Count First to CUT H	1	8	13	135	1,359	13,600	136,018
Count CUT L to Last	0	0	74	737	7,365	73,641	736,400
Total	1	8	87	872	8,724	87,241	872,418

This algorithm performs a function that any person interested in election fraud would need. the reason is that it can be used to covertly tag fraudulent records without drawing attention to them. For instance,

Definition

An “algorithm” is any sequence of steps followed to achieve a certain goal. Even simple commands, such as “DIR” in MS-DOS can be considered algorithms, despite their simplicity. On that level, all ID number assignment methods can meet the definition of having been assigned algorithmically. However, when the word algorithm is used, it usually connotes a higher order of complexity than is found in the DIR command or the standard auto-increment method used to assign ID numbers.

In New York and New Jersey, the algorithms used to assign, map, or manipulate ID numbers are complex. For that reason, they warrant special attention. For this paper, the issue isn’t whether “algorithms” were used to assign or modify Ohio voter roll ID numbers. Literally, they were. The real issue is whether the algorithms used were unnecessarily complex, performed hidden or inexplicable tasks, or exhibit any unusual characteristics.

Data source

This study began with a version of the Ohio Voter rolls that was generated shortly after 10/29/2020. That version was used to generate scatterplots of CID and State ID (SID) numbers. Another version was downloaded on 8/22/2024 and was used to make comparisons with the earlier database, to confirm that previously generated scatterplots remain consistent over time, and for analysis of the most current data.

The 2020 database contains 8,071,294 records. The 2024 database contains 7,995,785 records. The reduced number of records in the more recent database may reflect Ohio’s recent efforts to purge their rolls of inactive or disqualified voters.

In New York, the algorithms appear to have been introduced at about the same time as HAVA regulations were implemented in that state (approximately June, 2007). The same is true of North Carolina (2006), New Jersey (2007), and appears to be true of Ohio (January 2004), based on changes in CID numbers on or near that date across all counties.

The 2020 and 2024 Ohio databases lack purged records. The 2024 database uses "Active" and "Confirmation" for status designations. Notably, the "Confirmation" status in this context doesn't equate to "purged" status in New York, as individuals with confirmation status can still vote. This situation complicates algorithm analysis, suggesting that ineligible records are entirely removed from the database rather than being marked as purged.

For election and database integrity, Ohio’s method is more secure than New York’s, but it is also likely to create large gaps in the data that can impair discovery or understanding of algorithms used to assign voter ID numbers.

Irregulars

Before studying Ohio’s CID and State ID (SID) numbers for hidden relationships, it is useful to know if there are any suspicious records in the rolls. Without such records, there is little point to using a complicated system to manage them, unless the goal is to introduce fraudulent records later. For that reason, I performed some superficial queries, similar to those performed in New York and New Jersey. Some suspicious records were found, but not in quantities suggestive of the widespread fraud suspected in New York.

Some of the numbers are undoubtedly false positives, and some may have legitimate explanations. If they are like New York, the explanations will apply to a minority of the records involved (usually between 0.05%-10.0%).

Suspicious records

Clones: 15,061 (2020 DB) 15,720 (2024 DB)

This is the total number of records with matching first/last names and birthdate. There will be some that are coincidental matches of the same common name and birthdate, but these are typically very small in number. In a state of Ohio's size, we would conservatively expect no more than 500 such matches, representing 250 unique name/birthdate pairs.

The actual number of records identified as possible clones is much larger than expectation for coincidence, leading to the likelihood that the majority are genuine cloned records. The number of unique name/birthdate pairs is about 7,500, which is enough to determine the outcome of close elections.

The query used to identify the Ohio records is identical to the query described in New York election law to identify potential duplicate registration applications. In New York, the query found almost 1.5 million cloned records. The difference between the numbers found in Ohio and New York can be partially explained by the lack of purged records in Ohio's system. Nearly one third of all records in NY are purged, thus increasing the size of the pool of records available for search.

Fictitious registration date, 1/1/1900: 68,983 (January 1st registration dates, various years: 221,210)

This date is a well-known stand-in for an unknown registration date. Several states admit that dates like these (January 1st in even-numbered years in the remote past) are false. This seems to deal with the problem of unknown dates, yet it introduces false information into the database and is not consistent with competent database administration. Records like these should be corrected or deleted. Data validation tools would prevent a registration with a date like this from being made.

This false registration date can have a meaningful impact on election integrity because it is impossible to determine whether the voter was legally registered to vote at the time he cast votes in any prior election.

Missing Registration Date: 70,235

A missing registration date is material because it is impossible to determine whether the voter was qualified to vote in any given previous election. Missing a registration date makes the record incomplete, and incomplete records should not be processed.

Any off the shelf database program would prevent this type of entry by using data validation. It would also be able to highlight existing records that have date conflicts.

Registration spike in 2020

It is normal for registration numbers to spike in presidential election years. However, some spikes are so large that they merit investigation. In Ohio state, there were 877,292 new registrations in 2020. This is

10.87% of all registrations for all years combined. The 9 years comprised of 2012-2020 account for 50.84% of all registrations.

In 2020, 19.39% of all Cuyahoga County registrations were created. In the 9-year period from 2012-2020, 73.38% of all registrations were generated. Meaning, the preceding 111 years are responsible for less than 27% of Cuyahoga's total (Table 3). This does not take into account deleted records over time, but these numbers remain large regardless.

These statistics are interesting though not illegal on their face. They may simply reflect that Cuyahoga had the most successful voter registration campaign in Ohio history. Or, as in New York, it may reflect the introduction of many false records in that year.

Table 3 Cuyahoga County registrations by year compared to Ohio state

OH State	Years	Count records	Percent total records	Total records		Cuyahoga	Years	Count records	Percent total records	Total records
	2012	346,159	4.29%	8,071,294			2012	38,061	4.28%	888,257
	2013	225,836	2.80%				2013	26,204	2.95%	
	2014	283,837	3.52%				2014	38,227	4.30%	
	2015	427,396	5.30%				2015	53,515	6.02%	
	2016	681,512	8.44%				2016	105,690	11.90%	
	2017	281,824	3.49%				2017	40,899	4.60%	
	2018	558,148	6.92%				2018	104,807	11.80%	
	2019	421,797	5.23%				2019	72,180	8.13%	
	2020	877,292	10.87%				2020	172,201	19.39%	
			50.84%						73.38%	

The extreme registration spike in Cuyahoga County 2020 (19.39%) compared to Ohio's state-wide figure (10.87%) raises concerns about data accuracy.

A comparison of the 2020 and 2024 databases reveals a more disturbing statistic: the number of voters added in 2020, as recorded in both databases, implies that more records were removed from the database than appears possible unless many of those records were suspicious.

According to the 2020 database, Cuyahoga County registered 172,201 voters between 1/1/2020 and 10/29/2020. The 2024 database shows that 96,397 voters were registered in that same time frame. The difference between these numbers, 75,804, is 44.02% of the original batch of records. Nearly half of the new registrations made in the beginning of 2020 were missing from the database less than four years later.

This is peculiar because in Ohio, it takes a minimum of four years to delete records for inactivity. There are other reasons a record might be deleted, but these strain credulity. For instance, death or loss of civil rights due to court order. Both affect a minority of registrations, and in the case of death, are more likely to affect much older registrations than these, too new to be deleted for inactivity even if never used.

Another possibility is that the voters moved to other counties, but this would not only suggest a newsworthy mass migration, but also a larger than normal percentage of civic-minded individuals who notified their local BOE of the move.

CID and SID numbers

SID numbers are unique for all records, but CID numbers are not. Out of 7,995,785 records, 4,563,895 share a number with at least one other county. Some ID numbers, like 35, are found in 15 counties (Table 4).

Table 4 First 10 CID numbers, 10 Ohio counties

Franklin	Fulton	Gallia	Geauga	Greene	Guernsey	Hamilton	Hancock	Hardin	Harrison
1	1	1	3	4	17	5	701	2	1,002
3	1,004	2	8	6	32	9	702	400,001	1,003
12	1,005	5	9	8	34	15	711	900,001	1,004
13	3,100,001	6	12	9	35	27	722	3,700,001	1,006
15	7,000,002	7	15	17	37	28	728	3,700,002	10,027
16	7,600,002	9	16	18	44	33	734	3,800,001	10,029
18	7,600,008	12	17	20	50	53	741	5,600,002	10,033
21	7,700,001	14	19	26	51	55	744	6,000,001	10,037
22	7,700,003	18	23	28	67	56	745	6,100,002	10,041
31	7,800,001	24	25	31	69	62	750	6,700,001	10,058

The widespread sharing of CID numbers across counties rules out the possibility that unusually large gaps between numbers are due to efforts to avoid overlap between counties. For example, the lowest CID in Montgomery is 272, while in other counties, the lowest is 1. This discrepancy cannot be explained by reserving lower numbers for other counties, as we know numbers are freely shared. Therefore, the implication is that either 271 lower number records were deleted in Montgomery, or its numbering intentionally started at or near 272. This raises an important question: why does Montgomery's numbering deviate from the pattern seen in other counties?

Scatterplots

A scatterplot is a comparison of two values on the X and Y axes of a chart, to determine relationships between the values represented. Scatterplots for all 88 of Ohio's counties were generated as a way to quickly isolate counties of interest. Each plot maps County ID (CID) numbers to the X-Axis, and State ID (SID) numbers to the Y-Axis.

Many counties appear normal. That is, numbers appear to have been assigned sequentially. This is evident by a fairly even slope from zero (bottom left) to the maximum SID and CID number (upper right). Fairfield County is an example of a "standard" pattern (Figure 1). There are a few outliers, possibly voters who moved to Fairfield from another county, bringing their previous SID number with them.

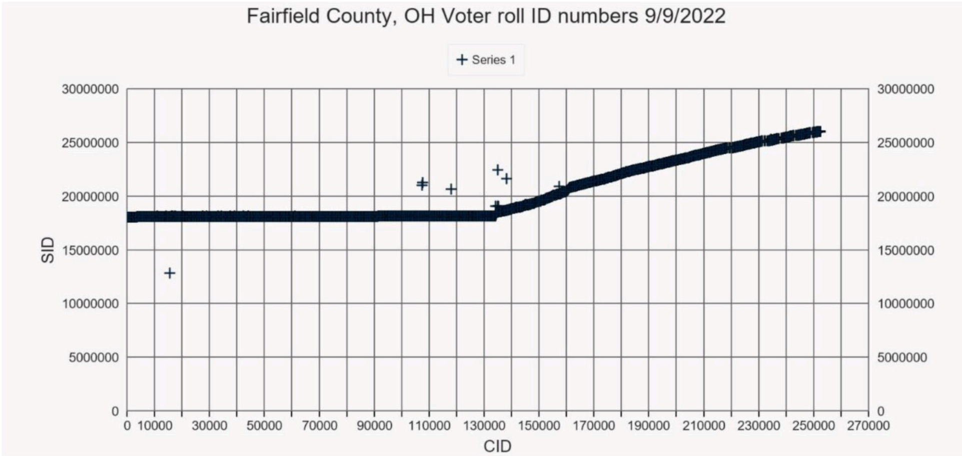

Figure 1 Fairfield County Scatterplot shows normal distribution of ID numbers

In contrast, Lucas County's plot does not show a regular or expected distribution of numbers (Figure 2).

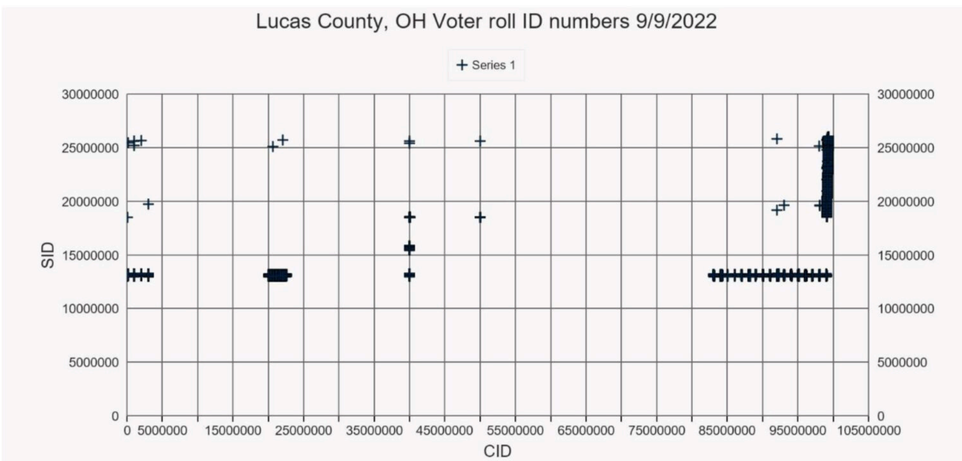

Figure 2 Lucas County scatterplot shows unexpected and irregular distribution of ID numbers

Lucas County's scatterplot, like some other similar plots, differentiate Lucas from the Standard plot and most Ohio counties. It tells us that the numbers are not assigned sequentially or chronologically because the full range of available SID numbers are used for each of several distinct groups of CID numbers. If this had been distributed normally, high SID values would only appear in the upper right quadrant of the plot. This is discussed in more detail in the Number space mapping section.

Franklin County's scatterplot, when filtered to include only pre-HAVA registration dates (about 1/2004), has a distinctive pattern (Figure 3). The pattern is of 23 horizontal bands of numbers, interrupted by spans without numbers. In total, each span and band combination are constructed of a band of

1,000,000 CID numbers followed by a span of 9,000,000 unassigned numbers, 10,000,000 in total. In each 1,000,000 number band, only a few of the available numbers are assigned, but the numbers assigned for this group only appear within this band, representing about 10% of the available number space. Each CID band is correlated with registration years from 1977-1999 with the exception of outliers, which constitute 8.05% of the total.

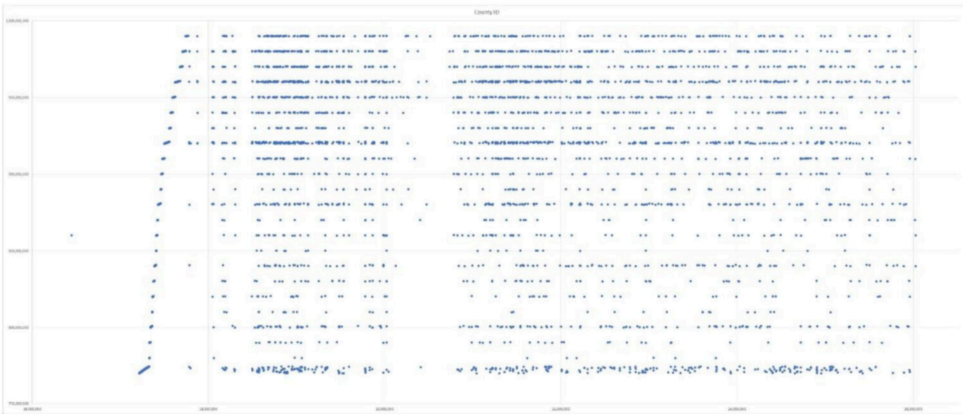

Figure 3 Scatterplot, Franklin County, Registration Dates up to January 2004. CID on Y axis, SID on Y.

This type of structure, where bands of numbers are segregated by registration year, is found in several Ohio counties. By allocating ten million numbers per year, with only a small fraction actually used, Ohio's number space is being consumed inefficiently. However, given the vast available range (over a billion numbers), exhaustion of the number space is unlikely to occur any time soon. This system, though inefficient, provides a clear chronological structure to the CID assignments across these counties.

Registration Dates

Registration dates and ID numbers were compared in Clinton County for correlations. Clinton was chosen because it had the lowest SID number (10,000,000) and the first long contiguous set of same county registrations (Table 5).

In Clinton County, SID numbers do not correlate with RegDates, indicating they were not assigned chronologically. CID numbers in the same county line up well with RegDates, with some exceptions, indicating that CID numbers were assigned chronologically, but that in some cases (possibly due to a voter moving) the RegDates are more recent than the CID number suggests.

Table 5 Registration date correlations, Clinton County, OH

Clinton County	**RegDate**	**CID**	**SID**	**RID**
RegDate	NA	partial	no	no
CID	partial	NA	no	no
SID	no	no	NA	no
RID	no	no	no	NA

As in Clinton County, Franklin County CID numbers are divided into blocks of approximately ten million numbers each. Within each block, most of the available numbers are assigned to the same year. (Table 6).

Table 6 Franklin County CID number distribution by Registration Year

CID Group	Year	CID MIN	CID MAX	Range	Records	Pct Range	Matched years	Outliers	Pct Outliers
3	1977	770,010,067	774,634,448	4,624,382	46,451	1.00%	45,268	1,183	2.55%
4	1978	780,010,086	780,213,963	203,878	1,946	0.95%	1,888	58	2.98%
5	1979	790,010,135	790,555,832	545,698	5,673	1.04%	5,484	189	3.33%
6	**1980**	**800,010,175**	**801,028,477**	**1,018,303**	**11,617**	**1.14%**	**11,157**	**460**	**3.96%**
7	1981	810,010,135	810,184,435	174,301	2,149	1.23%	2,051	98	4.56%
8	1982	820,010,057	820,494,208	484,152	5,733	1.18%	5,489	244	4.26%
9	1983	830,010,106	830,511,180	501,075	6,495	1.30%	6,196	299	4.60%
10	**1984**	**840,010,001**	**840,861,435**	**851,435**	**9,654**	**1.13%**	**9,247**	**407**	**4.22%**
11	1985	850,000,157	850,203,198	203,042	2,243	1.10%	2,114	129	5.75%
12	1986	860,010,023	860,365,050	355,028	5,201	1.46%	4,921	280	5.38%
13	1987	870,010,034	870,249,118	239,085	3,755	1.57%	3,551	204	5.43%
14	**1988**	**880,010,193**	**880,903,198**	**893,006**	**15,037**	**1.68%**	**14,174**	**863**	**5.74%**
15	1989	890,010,110	890,240,205	230,096	4,040	1.76%	3,787	253	6.26%
16	1990	900,010,088	900,384,778	374,691	7,407	1.98%	6,886	521	7.03%
17	1991	910,010,048	910,457,476	447,429	9,406	2.10%	8,414	992	10.55%
18	**1992**	**920,010,008**	**921,174,020**	**1,164,013**	**23,489**	**2.02%**	**21,226**	**2,263**	**9.63%**
19	1993	930,010,086	930,282,361	272,276	5,376	1.97%	4,705	671	12.48%
20	1994	940,010,038	940,408,538	398,501	8,621	2.16%	7,629	992	11.51%
21	1995	950,010,006	950,555,401	545,396	11,902	2.18%	10,333	1,569	13.18%
22	**1996**	**960,010,009**	**960,837,906**	**827,898**	**19,802**	**2.39%**	**17,247**	**2,555**	**12.90%**
23	1997	970,010,095	970,451,547	441,453	10,302	2.33%	8,720	1,582	15.36%

24	1998	980,010,071	980,488,826	478,756	11,930	2.49%	10,143	1,787	14.98%
25	1999	990,010,015	990,390,330	380,316	9,802	2.58%	8,238	1,564	15.96%
Totals					238,031		218,868	19,163	8.05%

This is presented as information only, to explain the banding found in some scatterplots. Otherwise, the pattern created by this characteristic of the CID numbers would remain subject to speculation.

SID Sort non-chronological

The lowest SID number in Ohio is 10,000,000, in county 14 (Clinton). It is mapped to CID 7,400,076, and has a RegDate of 4/8/1974. There are many earlier RegDates, but no lower SID numbers. This indicates that SID numbers are not assigned in chronological order. It also indicates that choice of CID is not chronological, but based on previously assigned numbers.

Non-chronologically assigned numbers are unusual if this is the rule rather than the exception. In Ohio, there are long strings of numbers that are chronological, interrupted by strings that aren't.

SID sort alphabetization

The SID sort is interesting because it largely correlates with an unusual type of alphabetization in 57 counties. If sorted by SID, names attached to those records appear at first glance to be alphabetically sorted by last name, first name, then middle name. On closer examination, the "alphabetization" is not alphabetized at all. Instead, the names are sorted into groups based on initials, not full names.

Below is the first such group of names. You will see the second and later characters are not taken into account for the purpose of alphabetizing the list. I am unaware of any software tool that will do this. It would be simple to make one, but that begs the question why. A list sorted like this is inherently difficult to navigate, making it inferior to default alphabetization methods.

***Abner**, Aber, Abirached, Abirached, Ables, Abner, Accoo, Accoo, Accoo, **Abt**, **Abt**, **Achtermann**, Achor, Achor, Achor, **Abner**, Achor, Achor, Achor, Achor, **Achtermann**, Achor, **Acuff**, **Adam**, Achtermann, **Ackels**, Ackerman, **Achtermann**, Ackerman, **Acuff**, Ackerman, Ackerman, Acuff, Acuff, Adams, Adams, Adams*

Clinton's records appear to have been generated in multiple pseudo-alphabetized batches, interspersed with records not sorted by name. The first group of SID numbers when sorted by SID is county 14 (Clinton), which uses SID numbers 10,000,000-10,021,637 (Abner-Zurface). After this, the next batch of Clinton records have non-alphabetized names for SID 10,021,640-10,021,714. It then returns to pseudo-alphabetical starting with the name Palmateer and SID 10,021,723 through 10,023,025 (Reiley). After this, the records switch to county 49.

This pseudo alphabetization resembles a loose hand sort rather than anything done in a computer. However, even that is unusual. Proper alphabetization techniques have been in use since long before the computer age. The significance of this finding is unknown but it bears further investigation. It would be surprising if a covert tracking algorithm manipulated names because they are not unique, unlike ID numbers. However, sorting like this could add unique properties to the list if it was deterministic.

Gap analysis

Gap analysis compares two values by subtracting one from the other to determine the difference. This method was used in New York, where it revealed the "Spiral" algorithm. It did this by exposing a Repunit-based pattern built into SID numbers. Adjacent numbers had regularly spaced gaps of Repunit numbers like 1,111, 111, and 11. Gap analysis is subjective because added or deleted records between existing numbers can change gap values. In New York, deleted records were reflected in gap values. If one record was missing and the previous value was 1,111, then the next gap would be 2,222 instead of 1,111 due to the missing record.

An initial gap analysis of Ohio's 88 counties revealed that many CID gaps appeared related to each other in at least three counties: Franklin, Lucas, and Montgomery. For instance, in Lucas County, these are the first 10 gaps in sequential order: 56, 8, 8, 16, 56, 8, 40, 32, 24, 16. All of these numbers are divisible by 8. One possible explanation is that the CID numbers were assigned in increments of 8. Later, records were deleted as people died or moved out of state, which caused gaps that are a multiple of 8 instead of 8, as in New York.

The explanation doesn't agree with the data in Ohio. This is because there aren't enough records in any of these counties to cover the range of numbers they use with an increment by 8 numbering system. The first 8,784 records in Lucas cover a span of 1,038,800 numbers. The most they could cover with an 8 increment is 70,272 (8,784*8). This indicates that the span was covered using an algorithm designed to spread out the numbers.

To explore this further, a gap frequency analysis was performed.

Frequency analysis

Gap Values

A frequency analysis can be used to determine if actual frequencies differ from expectation and by how much. In Lucas, for instance, the cluster of numbers divisible by 8 would be unusual if it continued throughout Lucas County's CID numbers. A way to check is to count how many times those numbers appear. This is the gap value frequency.

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

Table 9 Lucas County Gap frequency sub-pattern

Position	CID Gap Value	Lucas Frequency	Rank 1-100	Rank within 8
1	**16**	**14,121**	**5**	**1**
2	17	23	66	6
3	18	44	61	4
4	19	1,167	24	2
5	20	2	81	7
6	21	0	95	8
7	22	848	31	3
8	23	38	62	5
1	**24**	**9,415**	**6**	**1**
2	25	4	77	6
3	26	62	58	4
4	27	1,824	15	2
5	28	3	79	7
6	29	1	86	8
7	30	1,221	23	3
8	31	15	69	5
1	**32**	**6,571**	**8**	**1**
2	33	8	73	5
3	34	79	55	4
4	35	1,605	16	2
5	36	6	76	6
6	37	0	95	8
7	38	1,023	27	3
8	39	1	86	7

Franklin County has gap frequency spikes similar to those found in Lucas County. They are also based on the number 8, but not for the same reason and with built-in offsets. In Lucas, there appears to be a secondary and tertiary pattern within the main pattern so that the highest, second highest, and third highest frequency values in each block of 8 gap values appears in the first position, as if overlapped. In Franklin County, the offsets are stretched out linearly.

For instance, the highest value gaps are at: 8,16, 24, and then skips to 27, 35 (+8), 43 (+8), 51 (+8), 59 (+8), and then another offset to 65 (offset +6) after which it reverts to +8. There are a few other offsets, but they all follow this pattern.

The dominance of the number 8 and its multiples in CID gap frequencies is easily viewed in a scatterplot. To show the complexity of the pattern, an overlay was added to the following illustration (Figure 4). This illustration reveals that gap frequencies are controlled to create a highly specific waveform pattern. A visual inspection of the plot makes it clear that there are curves delineated by gap frequencies spaced 8 numbers apart. There are 8 of these curves with different start points. Each ascends, then descends at a predictable rate, but the peaks for each curve are shifted relative to each other.

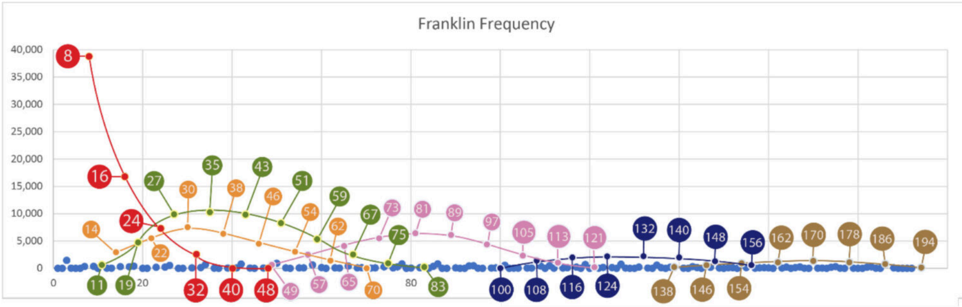

Figure 4 Franklin County, OH CID Gap Frequency curves based on Base-8 numbering

Based on the Figure 4 scatterplot, Franklin County's gap frequencies were segregated into separate curves. Each curve had a different starting point, then continued with a +8 modification. For instance, Group A starts with a gap frequency of 1, then adds 8 for the next frequency of 9, then adds 8 again to become 17, and so on. Group B starts with the number 2, Group C on 3, and continues in this fashion until Group H, which starts on 8. In this way, all gap values are accounted for as belonging to one of these 8 groups, and lie on the curves generated by them.

A table of these values, divided by strand and gap value, reveals a striking cyclic pattern (Table 10).

Table 10 Franklin County CID Gap value frequencies divided by group membership

Group A Gap	Group A Frequency	Group B Gap	Group B Frequency	Group C Gap	Group C Frequency	Group D Gap	Group D Frequency	Group E Gap	Group E Frequency	Group F Gap	Group F Frequency	Group G Gap	Group G Frequency	Group H Gap	Group H Frequency
1	0	2	0	3	1456	4	29	5	0	6	0	7	353	8	38771
9	377	10	0	11	623	12	41	13	34	14	2883	15	314	16	16865
17	285	18	414	19	4800	20	0	21	22	22	5423	23	213	24	7347
25	235	26	538	27	9930	28	19	29	0	30	7578	31	84	32	2535
33	160	34	670	35	10234	36	33	37	0	38	6296	39	71	40	0
41	99	42	801	43	9796	44	56	45	0	46	4545	47	79	48	72
49	723	50	935	51	8295	52	82	53	0	54	3005	55	87	56	87
57	2552	58	639	59	5303	60	146	61	0	62	1401	63	84	64	67
65	4158	66	429	67	2528	68	100	69	130	70	0	71	73	72	84
73	5427	74	293	75	1003	76	97	77	303	78	825	79	7	80	77
81	6491	82	110	83	256	84	47	85	388	86	812	87	23	88	0
89	6132	90	83	91	8	92	22	93	464	94	501	95	33	96	0
97	4373	98	73	99	83	100	8	101	576	102	2	103	35	104	43
105	2340	106	61	107	76	108	1389	109	441	110	13	111	498	112	39
113	1051	114	60	115	131	116	1978	117	230	118	5	119	715	120	18
121	280	122	39	123	171	124	2221	125	148	126	4	127	754	128	69
129	14	130	3	131	240	132	2211	133	81	134	3	135	594	136	132
137	12	138	204	139	257	140	1977	141	37	142	4	143	332	144	144
145	17	146	515	147	195	148	1177	149	62	150	0	151	128	152	137
153	16	154	795	155	140	156	589	157	47	158	84	159	130	160	101
161	15	162	1006	163	75	164	228	165	25	166	139	167	274	168	53
169	0	170	1346	171	34	172	66	173	20	174	196	175	322	176	35
177	0	178	1011	179	27	180	0	181	15	182	194	183	288	184	33

The voter identification numbers in Lucas County appear to be assigned using a system based on modular arithmetic. Numbers are grouped into discrete sets based on their modulus when divided by 8, creating several parallel sequences each with its own modulus-based pattern. This structure is the basis for calling the algorithm that produced it, "Modulus 8."

A modulo (Mod) value for a number is that number divided by the modulus, which in this case is 8. The Mod value returned from the calculation is the remainder. The numbers 8, 800, and 1600 have identical

Mod/8 values, 0, because they are all evenly divisible by 8. For other numbers, the mod ranges from 1-7. For instance, the mod of 33 is 1 because 33 divided by 8 is 32 with a remainder of 1, or Mod 1. The number 39 has a Mod of 7 because it is 39 -(8*4) =7.

Franklin CID number groups were reorganized by modulus remainder values 0-7, corresponding to the offset used for the first value of each "Strand". The modulus was calculated based on the absolute CID number. Gaps are based on the relative position of CID numbers in a full list of CID numbers sorted by CID. Strands of these numbers were extracted based on modulo value, including the relative gap values, and frequencies calculated. These were then made into a separate graph, to compare curve structure of gap frequencies for the first 74 gap values (Figure 5). The resulting graph is similar to the graph made of gap frequencies, a relative measure, though this chart is tied to the Mod values of CID numbers, an absolute measure.

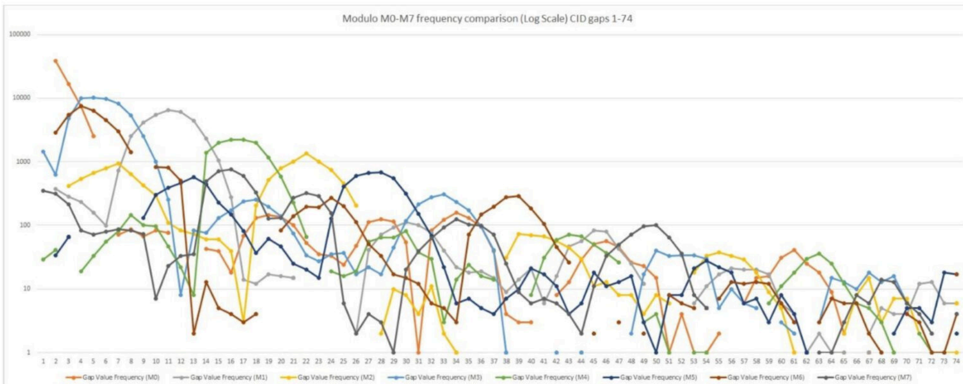

Figure 5 Franklin County, OH Strand frequencies, gap values 1-74 (0 values missing)

This graph shows clear non-random cyclic structure. The curves are not in modulo order. Instead, they follow a mathematical sequence starting with M3, followed by M1, M4, M2, M5, M3, M6, M4, M7, M5, etc. The numbers follow the pattern -2, +3, -2, +3, etc. Since -2+3=1, each iteration advances the modulo value by 1 and cause strand curve peaks to be 8 gap values apart.

Lucas County utilizes the Mod values differently. In Lucas, groups of numbers are organized by Mod. For instance, CID numbers 168 through 1,038,968 (n=8,784) are all Mod 0. The next 70,411 records are not grouped by Mod. After that, the next 13 groups (n=49,238) each use the same Mod, in this order: 5, 0, 4, 3, 5, 0, 3, 0, 1, 4, 6, 3, 0 (Table 11). In this way, Lucas County has effectively created a way to mark these records as distinct from others. This is called a hidden attribute. In this case, the hidden attribute is both the Mod value and the fact that the numbers are clustered into groups based on Mod.

Table 11 Lucas County Mod breakdown

Full List CID Group	Full List Mod Order	MOD	Row MIN	Row MAX	Records	CID MIN	CID MAX
1	Mod 0 to Row 8785	0	1	8,785	8,784	168	1,038,968
2	No disc. Order to Row 79197	NONE	8,786	79,197	70,411	1,038,987	85,049,172
3	Mod 5 to Row 79675	5	79,198	79,675	477	85,049,229	85,075,357
4	Mod 0 to Row 82438	0	79,676	82,438	2,762	86,000,016	87,072,024
5	Mod 4 to Row 88023	4	82,439	88,023	5,584	88,000,036	88,243,612
6	Mod 3 to Row 89480	3	88,024	89,480	1,456	89,000,011	89,058,699
7	Mod 5 to Row 95505	5	89,481	95,505	6,024	90,000,045	91,090,973
8	Mod 0 to Row 103604	0	95,506	103,604	8,098	92,000,040	92,287,096
9	Mod 3 to Row 106017	3	103,605	106,017	2,412	93,000,091	93,096,755
10	Mod 0 to Row 109130	0	106,018	109,130	3,112	94,000,008	95,016,888
11	Mod 1 to Row 112503	1	109,131	112,503	3,372	95,016,905	95,139,225
12	Mod 4 to Row 119248	4	112,504	119,248	6,744	96,000,028	96,220,492
13	Mod 6 to Row 122593	6	119,249	122,593	3,344	97,000,038	97,113,350
14	Mod 3 to Row 125983	3	122,594	125,983	3,389	98,000,019	98,103,955
15	Mod 0 to Row 128448	0	125,984	128,448	2,464	99,000,024	99,072,680
16	Frag Cycle M0-7. Order to Row 130252	Fragment	128,449	130,252	1,803	99,072,683	99,077,623
17	Cycle Mod 0-7 to Row 298955	0-7	130,253	298,955	168,702	99,077,626	99,366,553

This is not trivial. It requires at least some amount of planning to segregate numbers like this, and in so doing, to create information that is inaccessible via the normal user interface for databases like Ohio's voter rolls.

Montgomery County

A chi-square analysis was performed on the gap frequencies in Montgomery County's CID system, categorizing gaps by their remainder when divided by 8 (mod 8). The analysis yielded a chi-square value of 410,038.67 with 7 degrees of freedom. This result is statistically significant at $p < 0.001$, far exceeding the critical value of 24.32. This provides robust statistical evidence that the distribution of gaps is not random and follows a specific pattern. The data strongly suggests an algorithmic approach to CID assignment, with certain gap patterns being significantly more prevalent than would be expected by chance or natural record deletion processes (Table 12).

Table 12 Montgomery County gap frequencies, organized by Modulo and Rank

1 Mod 8 Rank	2 Mod 8 Rank	3 Mod 8 Rank	4 Mod 8 Rank	5 Mod 8 Rank	6 Mod 8 Rank	7 Mod 8 Rank	8 Mod 8 Rank
7	5	2	4	7	7	3	1
3	8	4	6	7	2	5	1
5	4	3	8	7	2	6	1
5	4	1	7	8	2	6	3
4	3	1	6	7.5	2	5	7.5
6	3	1	7	8	2	4	5
4	3	1	7	8	2	6	5
2	4	1	7	8	3	5	6
1	3	2	5	7	8	6	4
1	5	2	6	4	3	7	8
1	6	4	5	3	2	7	8
1	7.5	6	5	3	2	4	7.5
1	3	4	8	2	7	5	6
1	6	5	2	4	8	3	7
2	6	5	1	4	7	3	8
4	5	3	1	6	8	2	7
6	8	3	1	5	7	2	4
6	5	4	1	7	8	2	3
7	2	4	1	6	8	5	3
8	1	4	2	5	3	6	7
8	1	5	3	6	4	2	7
8	1	7	4	5	3	2	6
7	1	4	8	5	2	3	6
8	1	4	5	6	2	3	7
8	1	4	5	2	3	7	6
8	2	4.5	4.5	1	6	7	3
7	8	4.5	3	1	4.5	6	2
4	7	8	3	1	5	6	2
2	6	4	3	1	8	7	5
2	7	3	5	1	4	6	8
3	8	1	7	2	4	5	6
5	8	1	7	4	6	3	2
5	7	1	8	6	4	3	2
6	7	1	4	5	8	3	2
7	8	1	4	6	5	2	3
7	8	3	4	5	1	2	6
5	8	6	3	4	1	2	7
2.5	6	7	8	4.5	1	2.5	4.5
3	2	8	7	4.5	1	6	4.5
3	2	7	5.5	4	1	5.5	8
8	2	7	4	6	1	3	5
6	1	8	2	7	3	5	4
3	2	7	1	8	6	5	4
2	4	7	1	6	8	5	3
3	5	7	2	4	6	8	1
1	7	8	3	4	5	6	2
1	6	8	7	4	5	2	3
2	4.5	7	8	3	6	1	4.5
4	3	2	7	5.5	8	1	5.5
4	3	2	5	7	7	1	7

21

The gap frequency analysis reveals a striking anomaly at gap value 21. Out of 88 Ohio counties, 83 (94.3%) show zero occurrences of this gap, a pattern that stands out dramatically from surrounding gap values. For context, gap values 1-15 have non-zero frequencies across all counties. The number of counties with zero frequencies then increases gradually: 1 county for gap 16, 2 for gap 17, 5 for gap 18, 7 for gap 19, and 19 for gap 20. After the sharp spike to 83 counties for gap 21, the number drops back to 19 for gap 22 and continues to fluctuate thereafter (Table 13).

Table 13 Zero 21 gap values in 83 of 88 counties

CID Gap Value	Adams Fr	Allen Freq	Ashland Fi	Ashtabula	Athens Fr	Auglaize F	Belmont F	Brown Fre	Butler Fre	Carroll Fre	Champaig	Clark Freq	Clermont	Clinton Fr	Columbiar	Coshoctor	Crawford	Cuyahoga
14	27	194	52	58	188	13	20	24	846	22	36	71	408	24	42	11	25	2819
15	16	156	19	31	146	10	23	17	704	11	25	43	436	21	34	3	18	2199
16	16	149	21	36	137	3	17	16	566	15	19	27	364	9	15	5	12	1708
17	5	108	25	21	107	1	4	11	476	11	9	27	271	16	15	2	11	1404
18	15	113	28	19	90	1	6	8	375	6	16	20	236	7	10	1	10	1053
19	5	64	10	14	75	0	5	6	363	5	5	19	167	5	8	1	5	920
20	3	51	8	13	69	1	5	3	305	6	4	15	255	7	2	0	3	724
21	1	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0
22	1	44	10	5	49	1	2	3	210	0	2	7	149	1	2	0	4	474
23	2	27	6	5	39	0	4	0	184	0	3	12	115	4	2	0	0	393
24	2	33	5	5	55	0	4	0	148	0	1	7	123	1	1	0	3	320
25	1	26	2	3	34	1	3	0	137	0	2	3	163	0	0	1	3	265

This abrupt change for gap 21 is statistically significant. Assuming a random distribution of gaps, the probability of observing such an extreme deviation by chance is small. Using a binomial probability model and taking into account the decay rate in each county, the probability of 83 or more counties out of 88 showing zero frequency by chance is effectively zero.

The next highest count of counties with a zero frequency is 82, occurring at gap 95. However, this high count at 95 is less anomalous given the general trend of increasing zero frequencies for larger gaps. The isolated nature of the spike at gap 21 makes it particularly noteworthy.

Interestingly, the counties least likely to have a gap of 21 based on their number systems (Franklin and Montgomery, which prefer base 8 numbers) actually show non-zero frequencies for this gap. Conversely, counties where gap 21 should naturally occur more often show zero frequencies. This paradoxical distribution further underscores the non-random nature of this pattern.

While the data strongly suggests intentional avoidance of gap 21 across most counties, the mechanism behind this avoidance remains unclear, especially considering the possibility of gap creation through record deletion. This pattern warrants further investigation into the algorithms or processes governing voter ID assignment in Ohio counties.

Number space mapping

Scatterplots of Franklin, Lucas, and Montgomery CID and SID numbers reveal idiosyncratic but related structure based on modulus 8.

Lucas County's mapping pattern is distinguished by its isolated grouping of ID numbers in a number space that is much larger than needed for the size of the county, state, or any state in the United States of America. There are one hundred million numbers available in Lucas County's ID number format, to accommodate only 298,954 registrations.

If the numbers were assigned normally, we would expect to see a diagonal line that starts at the bottom left. This point represents the lowest CID and SID number. Then, it would ascend to the right, as sequential numbers are added. Instead, there are dense clusters of numbers in a "Block" pattern (Figure 6). These blocks exhibit a highly structured pattern that requires specific, predetermined rules for number assignment. Such a pattern is unlikely to emerge from random assignment or simple sequential numbering.

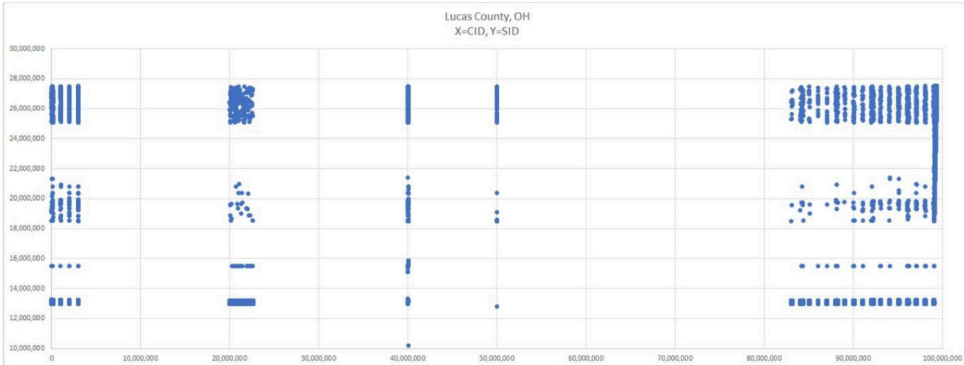

Figure 6 Lucas County, OH, scatterplot of CID and SID numbers, reveals block structure and Mod/8 substructure

The pattern reveals a structure where numbers are assigned only within specific, predetermined boundaries. These boundaries create a grid-like pattern with 20 sections (4 rows and 5 columns), where only a portion of each occupied section contains assigned numbers. The remaining space is completely void of numbers.

This pattern represents a far more complex undertaking than is typically necessary for the standard purpose of assigning unique serial ID numbers in a public database. While there might be unknown purposes that justify such complexity, from the perspective of basic voter registration ID assignment, this system appears unusually intricate.

By creating a diagram to position ID numbers in discrete groups, a programmer must personally set those boundaries and then ensure no duplicate numbers are generated. The second part of this task is not trivial, because the block shapes can only be created by repeatedly using the same range of CID numbers for successive narrow slices of SID numbers.

If, for instance, your CID range includes the numbers 1-100, this is like assigning the first 3 SID numbers to CID 1, 50, and 100. Then, for the second group of 3 SID numbers, assigning them to CID numbers 13, 54, and 89. If continued in this way, random number assignment could eventually lead to duplicated values. Without carefully designed rules to prevent duplication, this mapping system cannot work.

The block pattern observed in Lucas County's voter ID system raises concerns about potential covert tracking or encoding of hidden attributes. This structure could allow system designers to embed additional information not visible in standard data fields, creating a layer of data accessible only to those who implemented the system. Such a mechanism could be used for undisclosed tracking or flagging of specific voter records, bypassing normal transparency and oversight processes.

Lucas used at least 3 methods to assign CID numbers. The first is to group numbers by Mod. Second, they controlled numbers so that they covered a pre-defined value range. Third, SID and CID numbers were mapped to segregated blocks within available number space.

In combination, the block structure and differing uses of modulo based numbers cannot be accidental. It is built into the system used to assign ID numbers. It is different from most other Ohio counties, whose plots appear normal. The modulus arithmetic introduces a level of complexity well beyond what is necessary. The block diagram creates the potential for unintended error unless carefully maintained. Together, these qualities could be used nefariously if someone with administrative access chose to do so.

Possible explanations

The three Ohio counties of special interest, Franklin, Lucas, and Montgomery, used non-standard means to assign ID numbers to voters. This is of special interest because ID numbers are used to identify unique voters. Legitimate voter ID numbers are required to generate lawful ballots in elections.

There are a number of factors that could have led to the structure of data as described here.

- Shared database software/vendor: This is more likely than not, but doesn't explain the purpose behind using such a complex system
- Historical Data Migration: This is possible, particularly given the introduction of HAVA in early 2004. However, it doesn't explain the carefully designed aspects of CID number assignments, nor why it would have been done for older records but not new ones.
- Large County Size: The three counties of interest are among the largest in Ohio, but they are approximately the same size as a few other populous Ohio counties, such as Cuyahoga and Hamilton, neither of which uses the same methods used by Franklin, Lucas, and Montgomery. If county size was a factor in deciding to use a different system, it makes sense it would be used by all counties of similar size, but isn't.
- Early adopters of HAVA: While possible, HAVA applies to the entire state, not individual counties. Any solution designed to satisfy HAVA requirements should not differ between counties, regardless when HAVA was implemented in those counties. This also argues against local influence by certain county officials.
- Pilot/Test Program for data management: This would be an example of incompetence if it is the reason. Any such test should be performed with test data on a development database, not a production database, and then implemented statewide (not to only 3 counties) only if it was deemed worthy.
- Merged Databases: This could explain how the CID numbers we see today arrived in the current database. However, it does nothing to explain why they were generated this way.
- Error Correction: If this was the goal, it simultaneously introduces enough new ways to cause errors that could negate any potential benefit.
- Database Optimization: The CID number structure creates opacity in the data at the same time it creates hidden attributes. A fundamental rule of database management is, "Don't change the data." The patterns found in these three counties do effectively change the data by either adding attributes/information, or creating the opportunity to do it. Neither is consistent with database optimization, nor is it clear how these CID number assignment methods could optimize the database, if at all.

Conclusion

The goal of this paper was to determine whether the algorithms found in New York and New Jersey (and the tagging of ID numbers in Hawaii) are isolated or widespread phenomena. The findings recorded in this paper are enough to support the conclusion that algorithms associated with the generation of

County ID numbers have been found in Ohio's voter rolls. They are not the same algorithms as those found in New York or New Jersey, but they have shared characteristics.

The combination of the Modulus 8 and block mapping of ID numbers allow us to conclude the following:

- The existence of an unusual, unnecessarily complex voter ID system has been found in three populous Ohio counties
- The purpose of this system is questionable given its complexity and limited implementation
- A thorough investigation into the system's design, implementation, and current use is warranted
- The potential for this system to be used for voter data manipulation exists
- Weighed against other possibilities, a plausible explanation is that the algorithms used to create CID numbers in Franklin, Lucas, and Montgomery were designed for the purpose of covert data manipulation

References

(1993). National Voter Registration Act. 103-31. H.-H. A. S.-R. a. Administration. House of Representatives, Congress**:** 13.

Kaur, H. and J. Rani (2016). "A Survey on different techniques of steganography." MATEC Web of Conferences **57**: 02003.

NYBOE (2022). NY Comp. Codes R. & Regs. Tit. 9 §6217.5. 9. S. B. o. Elections. Albany.

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

Templ, M. and M. Sariyar (2022). "A systematic overview on methods to protect sensitive data provided for various analyses." International Journal of Information Security **21**(6): 1233-1246.

In view of the various methodological developments regarding the protection of sensitive data, especially with respect to privacy-preserving computation and federated learning, a conceptual categorization and comparison between various methods stemming from different fields is often desired. More concretely, it is important to provide guidance for the practice, which lacks an overview over suitable approaches for certain scenarios, whether it is differential privacy for interactive queries, k-anonymity methods and synthetic data generation for data publishing, or secure federated analysis for multiparty computation without sharing the data itself. Here, we provide an overview based on central criteria describing a context for privacy-preserving data handling, which allows informed decisions in view of the many alternatives. Besides guiding the practice, this categorization of concepts and methods is destined as a step towards a comprehensive ontology for anonymization. We emphasize throughout the paper that there is no panacea and that context matters.