The Caesar Cipher and Stacking the Deck in New York State Voter Rolls

A Paquette

New York Citizens' Audit

Email: paqart@gmail.com

Abstract: *Voters in New York State are identified by two identification numbers. This study has discovered strong evidence that both numbers have been algorithmically manipulated to produce steganographically concealed record attribute information. One of the several algorithms discovered has been solved. It first utilizes a mechanism nearly identical to the simple 'Caesar Cipher' to change the order of a group of ID numbers. Then, it interlaces them the way a deck of cards is arranged to create a 'stacked deck'. The algorithmic modifications create hidden structure within voter ID numbers. The structure can be used to covertly tag fraudulent records for later use.*

Keywords: *Information Warfare, Steganography, Cipher, Algorithm, Repunit, Voter Rolls*

Introduction

"Fraud detection comes into play once fraud prevention has failed" (Bolton & Hand 2002).

A review of publicly available data by the all-volunteer citizens' group New York Citizens Audit (NYCA) has found substantial evidence of fraud in multiple elections held in New York State (NYCA 2022). The types of fraud found by NYCA implicate a Non-Conventional Warfare (NCW) component. This is based on in-depth analysis of official state and county voter rolls, which appear to have been compromised at the state level. The database contains large numbers of fraudulent records accompanied by an algorithm that could be used to covertly access those records.

Fraudulent voter registrations can be generated innocently due to clerical or mechanical error. They can also be generated intentionally. The number and type of problematic registrations in New York State imply fraud more strongly than any kind of innocent error. Some are known to be the product of intentional fraud. This article is concerned with research that uncovered a well-hidden algorithm in New York's state voter rolls that can be used to covertly track fraudulent records. This algorithm has been implemented at the state level and is unambiguously intentional. There is no mention of this algorithm or its purpose in any of the voter roll-related contracts and other documents reviewed by NYCA. The algorithm satisfies a known need of any party engaged in an election fraud scheme: to covertly track and access fraudulent records. Knowledge of the algorithm can be used to accurately predict certain characteristics of voter roll records based on the algorithmically generated ID number alone. NYCA has been unable to determine how the algorithm was embedded in the state voter rolls or who was responsible for its design and implementation.

One advantage of information warfare is that identification of those responsible for cybercrimes is difficult. When identification is made, it is usually due to an informant coming forward who has direct knowledge of the persons or parties responsible (Atrews 2020). Without an informant, certain identification of an adversary is unlikely. This means that failure carries less risk, in addition to lower costs than CW. In 2014, the U.S. government was subjected to over 60,000 cyberattacks, just one of which successfully obtained almost 14 million social security numbers of government employees (Atrews 2020). It is not inconceivable that foreign or domestic adversaries have the ability to invade official government cyber territory and conduct IW operations there.

Findings from the research described in this paper demonstrate that New York State's voter rolls have been compromised. The rolls contain large numbers of fraudulent records and votes. This paper investigates an algorithm found in the rolls that has the capability to covertly modify records in a way that enables statewide election fraud.

Fraudulent records

The data uncovered by NYCA's research suggests that systemic election fraud is built into New York's electoral process. The current working hypothesis is that:

1. False voters were introduced into the voter rolls;
2. Records belonging to false voters were covertly tagged via an algorithm for easy retrieval when needed;
3. Absentee ballots were requested by false registrants;
4. Ballots and ballot envelopes were gathered at central collection points;
5. Fraudulently generated ballots were cast in fraudulently obtained ballot envelopes;
6. False voter records were updated to reflect false votes; and
7. After certification, false voter records were manipulated to disguise their purpose and history.

Of these 7 items, the following are known to have occurred:

- There are hundreds of thousands of illegally generated registrations in the official NYSBOE voter rolls. The exact number is unknown, but it is not less than about 338,000 for registrations active for the 2020 General Election (NYCA 2022). If other elections are included, the number of apparently illegal registrations jumps to between 1.2 and 2.4 million.

- 56.93% of all voter ID numbers were assigned based on the primary algorithm discussed in this paper. The algorithm allows a hidden attribute tag to be added.

- NYCA has recovered documents related to a fictitious identity with twenty-two registrations that requested multiple absentee ballots sent to the same rented mailbox. NYCA has identified other fictitious identities like this.

- Canvassing has uncovered cases where false votes were added to false registrations or genuine votes were erased.

- A comparison of four versions of the NYSBOE voter roll database created over a thirteen-month period shows hundreds of thousands of modifications to multiple fields belonging to the same voter ID numbers. Although there are valid reasons to update these fields, none of those reasons apply in these situations. For instance:

- In Greene County, a voter with the initials "C.S." has a DOB of 5/5/1925 in the 2021 database. In the 2022 database, his DOB is 8/18/1971. Both records have the same ID numbers, addresses, and RegDates. It is the same person, but birth dates are immutable. They do not change. Both records cannot be correct.
- A voter with initials "R.V." has three records in the 10/21 and the 12/21 databases. In the 2021 database, he has two SBOEID numbers. In the 2022 database, he has one. An examination of the records shows that one record, with a RegDate of 6/9/2021, was retroactively altered to change the SBOEID number. This is illegal because no voter is allowed more than one SBOEID number and changing this after the fact conceals the prior existence of an illegal record.
- A voter with initials "M.P." has two SBOEID numbers, one of which is illegal, and both have a vote recorded for the 2020 General Election. When M.P. was canvassed, she said she did not vote in that election, nor was she registered at the time of the election. The voter rolls confirm that she registered almost 3 weeks after the election, on 11/23/2020. Her voting history does not reflect her actual voting behavior.

It is unknown whether fraudulently gathered ballots were cast, but it is known that ballots were fraudulently requested and that votes were recorded as cast by voters associated with fraudulent SBOEID numbers. These two types of evidence indicate the probability that physical ballots were fraudulently cast.

An explanation proffered by some county officials consulted by NYCA for the existence of illegal registrations is that they are the product of innocent clerical error by incompetent employees. The algorithm found in the rolls argues against this. It is complex and precise. It is not the product of incompetence. This suggests election fraud rather than voter fraud. 'Election fraud' requires official access to election systems. It is distinct from 'voter fraud', which is committed by individual voters.

Algorithm overview

NYCA has discovered the presence of algorithms used to connect state (SBOEID) with county (CID) voter ID numbers in New York's 62 counties. Their structure can be used to covertly tag fraudulent records for later use. The presence of the algorithms was detected during an analysis of registration dates and their corresponding SBOEID numbers. Investigation revealed that CID and SBOEID numbers were linked by an algorithm that caused a sort of one ID number field to reveal an algorithmically produced pattern in the other. To make it more difficult, the 'sort and see' technique only works if the data is properly filtered. Filtering the data requires knowledge not normally available to users of the NYSBOE database.

The primary algorithm is found in 58 of 62 New York counties. NYCA has dubbed it 'The Spiral' because it wraps around itself in ever widening bands. The Spiral takes a sequence of CID numbers, translates them based on the algorithm, then assigns consecutive SBOEID numbers to the translated CID numbers. The effect of this obfuscates the presence of the algorithm and creates an invisible structure within ID numbers (**Table 1**). All CID and SBOEID numbers are translated in a similar fashion, making them predictable to anyone with knowledge of the algorithm but invisible to everyone else.

Record	Record Gap	CID Sort Row	CID	CID gap	SBOEID	SBOEID Gap	SBOEID Sort Row	Row Xform
4,061,739	0	18,465	100,136	0	22,530,891	0	307	18,158
4,061,850	111	18,466	100,137	1	22,531,002	111	418	18,048
4,061,961	111	18,467	100,138	1	22,531,113	111	530	17,937

Table 1: Algorithm-driven row transforms

The algorithmically imposed voter ID number structure creates a new but predictable relationship between their original (consecutive) order and the algorithm-imposed sort order. The pattern is more complex than hinted at above but remains predictable as long as the algorithm is well-understood. It is unrelated to other fields in the database and thus unlikely to have arisen naturally within the data.

Some potentially legitimate explanations for The Spiral's presence in the NYSBOE database are data privacy, search optimization, ease of use, and hacking prevention. For each explanation, The Spiral either does nothing or is worse than not using the algorithm.

The Spiral algorithm cannot protect Personal Identifying Information (PII) because the information attached to ID numbers is public. For the same reason, registration number guessing to find records in the database is unnecessary, making The Spiral superfluous. Hacking is unnecessary when all one has to do is send a request to the Board of Elections (BOE) for the voter rolls.

The Spiral reduces search efficiency by dramatically increasing path length between some records and reducing it equally dramatically between others. The savings on one side of the search terrain cannot be compensated on the other, unlike a well-balanced B-tree search (Sikdar 1992). It destroys the natural link between ID number sequences and registration dates that would aid users of the system to understand a record's position within the database. If that linkage were preserved, a user could estimate a record's age by looking at the registration number. The algorithm does one thing: it alters the structure of voter ID numbers.

The algorithmic sort order creates the appearance of compliance with public disclosure laws while concealing attribute information. The attribute information is uniquely available to keyholders, much as a card cheat has unique access to a straight flush in a stacked deck. Concealing information in plain sight, as was done in New York's voter rolls, is called 'steganography' (Kaur & Rani 2016). In combination with known fraudulent registration records, The Spiral algorithm presents the possibility that it has been inserted into voter roll software, or used to alter the NYSBOE voter roll database, for nefarious reasons. The Spiral can be used to identify fraudulent records quickly and covertly by repositioning records into key positions. Records of interest can then be extracted by software designed to recognize the algorithmically modified data structure.

Election Law

The Voter Registration Act of 1993 states that, "Each state shall maintain for at least 2 years and shall make available for public inspection…all records concerning the implementation of programs and activities conducted for ensuring the accuracy and currency of official lists of eligible voters" (National Voter Registration Act 1993). This sentence describes the scope and purpose of NYCA's investigation.

Each voter's record includes their names, dates of birth, residential addresses, party affiliations, voter histories, and other information. Each county varies in the number of data fields they record (or chose to supply to NYCA). Social security numbers (SSN) and driver's license ID numbers are the only fields that must be withheld from the public and they were withheld from NYCA (NY Election Law §3-220). All the data analyzed in this article is public and is derived from public sources.

New York uses an electronic voter registration list known as 'NYSVoter'. NYSVoter is maintained by the NYSBOE. The list, "shall maintain one record for each registered voter including the statewide unique identifier" (NY Election Law §6217.1) The same law provides County Boards of Elections (BOE) the "sole responsibility for adding, changing, canceling, or removing" voter records from the NYSVoter list (NY Election Law §6217.1). Any modifications made at the state level would violate this law.

According to state and federal election law, NYSVoter must provide county boards of elections the ability to query the statewide database. The query tool must allow sorting records by "county, election district, jurisdiction, birth date, and other information (*e.g.,* last name, first name, voter registration number, unique identification number, address order)" (NY Election Law §6217.12). It says nothing about sorting by a hidden algorithm key.

New York uses two ID numbers. One is called the "State Board of Elections ID" (SBOEID). The other ID is the "County VR Number" (CID). Voters are allowed one SBOEID number that "will remain with the voter for their voting life" (NY Election Law §6217.6). Excess SBOEID numbers are illegal under this law. Any voter who has two or more unique SBOEID numbers has been 'cloned'.

Any voter may be legally assigned multiple CID numbers if no two are simultaneously active. These are generated pursuant to a move from one county to another. NYCA has discovered cases where multiple CID numbers were generated without voter knowledge or a change of address.

SBOEID numbers use the format: "NY000000000012345678." The first two characters, "NY" and ten leading zeroes are identical in all SBOEID numbers found in the NYSBOE voter roll database. For analysis, SBOEID numbers were shortened to the last eight digits, called a 'Short ID'. This is sufficient to differentiate any two SBOEID numbers. This convention is followed in this article.

CID numbers vary between counties. Some are five digits, and some are nine digits. Some are alphanumeric; some are not.

Assignment of CID and SBOEID numbers

NYCA sent email requests to each of New York's 62 county Board of Elections (BOE) to ask how Voter ID numbers are assigned. They received 29 responses. Notably, none of New York's 10 most populous counties responded. The officials who responded hold the titles, "Deputy Commissioner", "Democratic Commissioner", and "Republican Commissioner".

All 29 commissioners stated that CID numbers are generated "automatically" by their "voter registration system" or by "software". NTS Data Systems was identified as the name of the software used by 12 counties. Fifteen counties stated that CID numbers are assigned sequentially or "simultaneous[ly]".

Essex County Democratic Deputy Commissioner Jen Fifield was the only respondent who stated that her county did not use NTS to manage their database. "We have our own in-house registration system," she wrote (Fifield 2022).

Thirteen of these answers lack any detail beyond the fact that CID numbers are assigned by software used by the CBOE. The remaining answers have enough detail to conflict with each other or findings uncovered by this research.

If each county's records are sorted by registration date (RegDate), CID numbers do not fall into a consistent sequential order. The same is true of SBOEID numbers. It is possible to find fairly long sequences within any given county's rolls, but the pattern is always broken multiple times in both directions. Dates and numbers ascend for dozens of entries, and then the dates drop backwards by year, as the ID numbers continue forward, then back again, and so on (**Table 2**)

	Jefferson County				
Reg Date Sort					
Reg Date	**CID**	**SBOEID**	**Reg Date Rank**	**CID Rank**	**SBOEID Rank**
1/1/1850	N920235	22,246,089	1	48,061	44,156
1/1/1850	N925948	22,251,736	2	53,065	49,725
01/07/1902	213946	22,202,061	3	30,739	230
09/16/1906	214307	22,202,247	4	30,907	416
04/07/1907	220571	22,205,310	5	33,661	3,477
CID Sort					
10/09/1962	10,224	22,207,470	635	568	5,629
10/10/1962	10,255	22,208,445	3,929	569	6,601
10/11/1962	10,262	22,207,581	672	570	5,739
10/11/1962	10,267	22,208,446	21,912	571	6,602
10/11/1962	10,277	22,207,692	713	572	5,850
10/11/1962	10,285	22,208,447	3,930	573	6,603
SBOEID Sort					
10/06/1962	007037	22,201,832	83	298	1
01/30/1996	213445	22,201,833	27,197	30,533	2
01/30/1996	213447	22,201,834	27,198	30,534	3
01/30/1996	213450	22,201,835	27,199	30,535	4
01/30/1996	213451	22,201,836	27,200	30,536	5
01/31/1996	213461	22,201,837	27,206	30,537	6
01/31/1996	213462	22,201,838	27,207	30,538	7
01/31/1996	213463	22,201,839	27,208	30,539	8
10/02/1971	026271	22,201,840	3,758	2,217	9

Table 2: Jefferson CID number samples, sorted by Reg Date, CID, and SBOEID numbers. Rank order for each criteria is different, regardless of sort method.

This research shows that the reason CID and SBOEID numbers are not sequential or random is that they are governed by the same algorithm. The algorithm uses CID numbers to force SBOEID numbers into a complex order inaccessible to normal users of the voter roll database. The algorithm order prevents sequential or random ID number assignment.

The fact that some counties use their own custom software to assign CID numbers, yet their roles are affected by The Spiral regardless, indicates that the modifications occur after the records leave county BOE custody. This may be relevant in the context of New York's election law §6217.6 if it is found that adding an algorithmically concealed attribute to SBOEID numbers constitutes alteration of records.

Methodology

NYCA began its investigation with a statistical analysis of voter turnout. The goal was to find aberrations from 'normal' data contained in the New York State Board of Elections (NYSBOE 2021) voter roll database. In high transaction volume industries, like banking, telecom, and insurance, statistical methods allow the high volume of potentially fraudulent transactions to be reduced to a manageable number for investigation (Becker, Volinsky & Wilks 2010).

Statistical fraud-detection methodologies can be effective but also suffer from several weaknesses. Digit-based tests, like Benford's Law, rely on an assumption of what a normal distribution of numbers should be in a fraud-free environment. If the assumption is false, the tests cannot reliably return usable results (Beber & Scacco 2012). This is analogous to NYCA's finding that, although voter turnout appeared to have been artificially manipulated based on voter age, there was no reliable baseline to compare it to. Another drawback is that it indicated unnaturally homogenous turnout proportions for the entire state without narrowing the scope of the investigation. Fraud detection via an Adaptive Exponentially-Weighted Moving Average (EWMA) works in high volume transactions, such as credit cards and phone calls (Becker, Volinsky & Wilks 2010) but is not suited to a low traffic environment, such as elections, where voters, no matter how numerous, cannot interact with the system more than twice a year.

The statistical tests run by NYCA did not narrow the scope of suspicious records or identify specific instances of fraud. To do that, the voter roll data was manually examined and compared with election law. As stated by Hand (2010), "Fraud detection is not something that can be pursued in the abstract. Understanding of and familiarity with the data is an essential key to effective detection". After violations of law were discovered, NYCA programmers wrote SQL queries to find more examples. Those efforts usually succeeded. It was during this process that two significant discoveries were made. The first was a large quantity of illegally generated registrations (NYCA 2022). Second was an algorithm that restructured voter ID numbers. The algorithm made it possible to covertly tag illegally generated records for later use. These two discoveries present the possibility that the New York Board of Elections (NYSBOE) operational security boundary has been compromised.

Findings

Filtering by county ranges

The eight-digit number format used for SBOEID numbers allows a possible 99,999,999 unique numbers. There are 20,765,242 (20.76%) numbers used within that range in the October 2021 database used for this research. The minimum value used is 03,306,104 and the maximum is 61,106,878. Between those two values, SBOEID numbers appear to be randomly distributed but they are not. After careful analysis, NYCA discovered one band of numbers, called 'In-Range' (IR), that were different from the rest. The band was difficult to find because it is sandwiched between two other bands, 'Out of Range High' (OOR H) and 'Out of Range Low' (OOR L). Each of 62 counties is assigned a range of SBOEID numbers within the IR band. Numbers in both OOR sections (partitions) are not segregated by county, as IR numbers are. The noise produced by OOR numbers obfuscates the presence of IR numbers. There are 11,822,181 (56.93%) SBOEID numbers in the IR partition (**Table 3**).

	Minimum	Maximum
Out of Range	40,481,162	99,999,999
In Range	8,502,559	40,481,161
Out of Range	0	8,502,558

Table 3: SBOEID (Short ID) ranges, IR, and OOR partitions

The NYSBOE assigns 'County Code' (CC) numbers to each county. The numbers are assigned alphabetically. Albany is CC# 01, Allegany is CC# 02, and so on through to Yates County, CC# 62. Despite this, the 62 partitions assigned to each county in the IR partition are not in alphabetical order.

Each county's SBOEID range is separated by a gap of two unassigned numbers. In every county, nearly all the available numbers have been assigned (89.95% overall, over 99% in each of 34 counties), leaving little room for new voters within the IR partition.

County	County Code	CRID	MIN SBOEID	MAX SBOEID	Gap to previous	MIN to MAX size	Used Numbers (NYBOE)	Percent used of available	Registered full range (NYBOE)
Out of range		1	0	8,502,558	0		127		
Schoharie	47	1.01	8,502,559	8,521,213	1	18,655	18,447	98.89%	21,134
Buffer 1		2	8,521,214	9,091,766	1		0		
Onondaga	34	2.01	9,091,767	9,382,492	1	290,726	290,015	99.76%	329,306
Schenectady	46	2.02	9,382,494	9,477,662	2	95,169	94,044	98.82%	109,164
Oswego	38	2.03	9,477,664	9,557,731	2	80,068	79,339	99.09%	83,022
Niagara	32	2.04	9,557,733	9,694,852	2	137,120	132,790	96.84%	150,686
Suffolk	52	2.05	9,694,854	10,584,725	2	889,872	882,440	99.16%	1,116,934
Essex	16	2.06	10,584,727	10,611,715	2	26,989	26,685	98.87%	27,222
Buffer 2		3	10,611,716	20,005,105	1		0		
Hamilton	21	3.01	20,005,106	20,010,209	1	5,104	5,054	99.02%	4,677
Columbia	11	3.02	20,010,211	20,054,800	2	44,590	43,484	97.52%	49,665
Franklin	17	3.03	20,054,802	20,082,320	2	27,519	27,314	99.26%	29,083
Warren	57	3.04	20,082,322	20,125,302	2	42,981	42,479	98.83%	48,505
Fulton	18	3.05	20,125,304	20,157,206	2	31,903	31,577	98.98%	35,632
Tioga	54	3.06	20,157,208	20,189,678	2	32,471	32,254	99.33%	35,581
Montgomery	29	3.07	20,189,680	20,221,054	2	31,375	31,080	99.06%	30,712
Seneca	49	3.08	20,221,056	20,241,573	2	20,518	20,309	98.98%	22,052
Madison	27	3.09	20,241,575	20,284,049	2	42,475	42,106	99.13%	45,868
Allegany	2	3.1	20,284,051	20,312,118	2	28,068	27,580	98.26%	27,588

Table 4: Partial list of IR counties, sorted by county specific SBOEID number ranges

The County Range ID (CRID) number found in **Table 4** allows filtering by IR partition numbers assigned to each county. The task of determining county ranges was complicated by voters who move from one county to another while retaining the SBOEID number of the county of origin. To find the county of origin boundaries, all SBOEID numbers had to be manually examined.

There is no known method within the NYSBOE or county BOE databases to discover the existence of IR and OOR partitions. Awareness of the partitions and knowledge of each county's boundaries are required to obtain a noise-free sample of the algorithm.

Records must be filtered to include only records with a RegDate earlier than 6/1/2007. Based on an analysis of RegDates (**Figure 1**) this appears to be the date when the algorithms were introduced. In most counties, IR records after this date are duplicates that create noise in the algorithm.

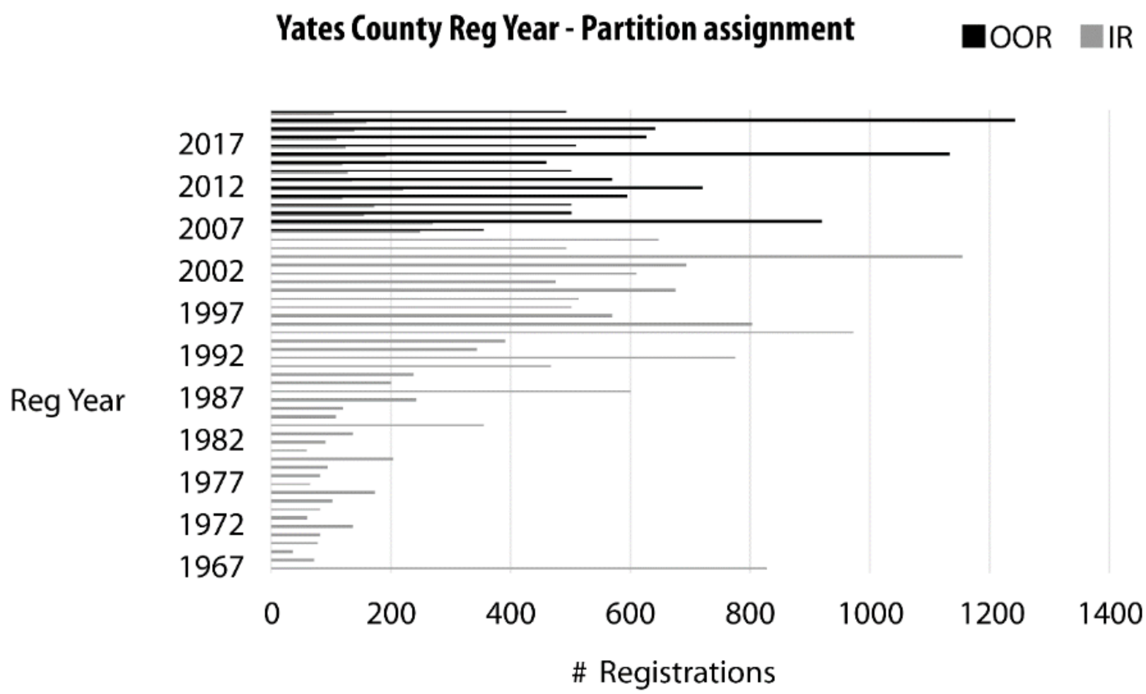

Figure 1: Yates County RegDates by year and partition assignment

Spiral Algorithm

After the data is filtered, it must be sorted. In **Table 5**, it can be seen that a CID sort creates an easily understood pattern in the gaps between SBOEID numbers. If each SBOEID number is subtracted from the next, the gaps between almost all of the numbers shown are either 1,111 or, at every tenth number, 1,112. A RegDate sort scrambles CID and SBOEID numbers.

If the records are sorted by SBOEID number, the algorithm appears in the CID column. In the group of twenty numbers in **Table 5**, there are three different types of numbers represented. The first has one digit, most have five digits, and two have four digits. Unlike the CID sort, where every tenth record is modified by adding one, as in 1,111+1=1,112, with an SBOEID sort, every eleventh record changes from a five-digit CID number to a four-digit CID number. This makes sense if the goal is to scramble voter ID numbers.

SBOEID sort			CID sort				RegDate sort		
RegDate	CID	Short ID	RegDate	CID	Short ID	Short ID Gap	RegDate	CID	Short ID
10/2/1971	1	20,284,051	10/02/1971	1	20,284,051	0	5/21/1911	3,554	20,299,837
3/21/1995	13,214	20,284,052	11/30/1992	2	20,286,829	2,778	2/5/1921	37,235	20,299,561
10/14/1967	13,220	20,284,053	09/15/1993	4	20,297,940	11,111	5/22/1922	20,699	20,289,226
5/29/1967	13,222	20,284,054	10/09/1984	7	20,306,285	8,345	1/21/1932	12,323	20,311,502
10/11/1984	13,223	20,284,055	11/18/1993	8	20,307,396	1,111	11/15/1933	2,152	20,289,659
5/29/1967	13,224	20,284,056	10/02/1971	9	20,308,507	1,111	1/17/1938	17,998	20,287,316
8/24/1989	13,225	20,284,057	10/13/1967	10	20,309,618	1,111	6/30/1938	14,962	20,285,212
8/24/1989	13,226	20,284,058	09/24/1987	15	20,310,729	1,111	10/11/1939	287	20,292,690
7/18/1967	1,340	20,284,059	04/19/1991	16	20,311,840	1,111	7/20/1940	4,172	20,304,403
10/10/1968	13,228	20,284,060	09/17/1977	17	20,284,884	-26,956	12/9/1940	6,896	20,307,756
10/10/1968	13,229	20,284,061	01/06/1994	19	20,285,995	1,111	6/16/1947	47	20,302,662
10/6/1972	13,236	20,284,062	01/04/1993	20	20,287,107	1,112	7/22/1949	15,069	20,285,294
10/5/1974	13,237	20,284,063	04/04/1968	23	20,288,218	1,111	5/20/1950	15,371	20,285,517
10/15/1983	13,238	20,284,064	04/04/1984	25	20,289,329	1,111	8/1/1950	327	20,295,356
7/18/1989	13,239	20,284,065	04/04/1984	26	20,290,440	1,111	10/28/1950	11,845	20,311,176
1/6/1992	13,240	20,284,066	06/20/1994	27	20,291,551	1,111	4/27/1952	1,434	20,284,725
1/6/1992	13,241	20,284,067	10/05/1976	31	20,292,662	1,111	5/24/1952	1,075	20,310,104
10/4/1975	13,244	20,284,068	10/05/1976	33	20,293,773	1,111	8/25/1953	7,325	20,308,049
10/4/1975	13,245	20,284,069	10/05/1974	36	20,294,884	1,111	10/6/1954	926	20,308,993
10/10/1972	1,341	20,284,070	10/05/1974	37	20,295,995	1,111			
			01/06/1992	38	20,297,106	1,111			
			01/06/1992	39	20,298,218	1,112			

Table 5: Sort methods compared by SBOEID, CID, and RegDate

Hackers sometimes try to enter a target system by guessing ID numbers. One way to prevent or impede such efforts is to randomize ID numbers. The reasons this does not apply to voter rolls are the voter rolls are public and the algorithm does not randomize numbers. Hacking is unnecessary and carries legal risk not associated with making a Freedom of Information Law (FOIL) request for voter rolls. The algorithm, if discovered, is highly predictable, the opposite of random.

Repunits

The constants of The Spiral pattern are numbers known as 'repunits'. A 'repunit' is a number of two or more digits composed of repeats of the number '1'. The numbers '11', '111', '1,111', '11,111', and '111,111' are all repunits (Francis 1988). In a CID sort, each group of 10 SBOEID numbers is bounded by a repunit that ends in 2. The numbers used to designate the last record in a block are 'End Tags'.

The algorithm is based on repunits and numbers related to repunits. The most common are:

- Full repunits, like '1,111'
- Repunit +1, like '1,112'
- Quarter (25%) repunits, like '278' (277.75)
- Three-quarter (75%) repunits, like '833' (833.25)

To understand how repunits are used, and the complexity of the overall pattern created by the algorithm, one must understand how it is constructed.

Structure

Caesar cipher

The Caesar cipher is a linear cipher. A linear cipher transforms each character of plaintext to create the encoded ciphertext. The Caesar cipher transforms each letter of the alphabet three positions to the right and then transforms the last three letters to the first three positions of the alphabet (Luciano & Prichett 1987). The transformations found in the voter rolls are similar to the Caesar cipher (**Figure 2**).

Figure 2: The 3 steps of the Caesar Cipher

To implement the Caesar cipher, The Spiral algorithm first separates a county's IR numbers into logical 'Strips' (**Figure 3**). Each strip is analogous to the alphabet sequence used by the Caesar cipher. These are series of continuous CID numbers with (usually) logical start and end points. The most commonly found strips are based on the number of digits or the alphabetical character(s) at the beginning of the CID number.

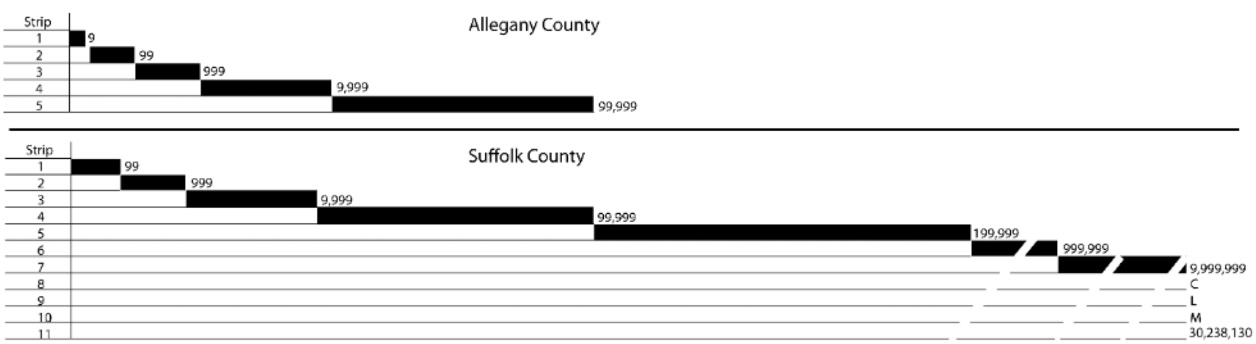

Figure 3: 'Strips' of CID numbers for Allegany and Suffolk counties (not to scale)

After the full range of a county's CID numbers are unequally divided into five or more strips, the Caesar cipher is applied. Each strip is cut near its beginning, creating two segments. Segment one contains the lowest numbers, and segment two contains the highest numbers. The last step repositions segment one to the end of segment two. This creates the pattern Strip1 Seg2, Strip1 Seg 1, Strip2 Seg 2, Strip 2 Seg 1….

Typically, the first and last strips have very few numbers, sometimes only one, as start and end points. The second and next to last strips are cut and transformed following the pattern established by the Caesar cipher. Remaining strips have a third step. A group of numbers from the repositioned Segment 1 and the end of Segment 2 are interlaced by introducing a group of low numbers into the high number range at regular intervals.

For example, in Allegany County, Strip 3 is comprised of numbers from 103 through 999. This strip was cut between CID numbers 145 and 146. Numbers 103-145 became Segment 1 and were moved to the end of segment 2, which starts with CID number 146. Then, the seven numbers between 103 and 114 (some numbers are missing) are interlaced every ten numbers with the last seventy numbers leading to 999 (**Table 6**). That portion of the strip is a 'sequence'.

	First CID (Start 2nd segment)	End continuous	Start alternating	Interval	Repeat	End 2nd segment	Start 1st segment	Last CID
Strip 01	1							9
Strip 02	17					99	10	16
Strip 03	146	911	103/913	10	7	999	115	145
Strip 04	1,340	9,516	1,001/9,518	10	30	9,997	1,047	1,339
Strip 05	13,214					91,126	10,001	13,213
	(black)	2nd strip segment						
	(white)	1st strip segment						
	(grey)	Alternating segment						

Table 6: Allegany County strip schematic

Deck stacking

The next step The Spiral uses to create number sequences is identical to the method used by card cheats to stack a deck of cards. Deck stacking is the act of pre-selecting a desired group of cards, such as a royal flush in poker, and then positioning those cards within a deck so that they are dealt to the desired player (**Figure 4**) (Clark 1986).

P1	P2	P3	P4	P5
3 ♥	2 ♦	A ♥	10 ♦	Q ♣
6 ♥	A ♣	K ♥	8 ♠	6 ♦
8 ♣	J ♠	Q ♥	9 ♣	5 ♦
K ♦	7 ♥	J ♥	8 ♥	5 ♠
10 ♦	2 ♠	10 ♥	3 ♣	7 ♥

P1	P2	P3	P4	P5
1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

Figure 4: Royal flush for Player 3 (P3) in a stacked deck (left) and the position of the cards in the deck (right)

The Spiral algorithm stacks the deck by interlacing numbers from different strips. Each strip is analogous to individual player's hands in the deck stacking example. Each number of each strip is spaced based on the repunit divisor for that strip. For instance, Strip 2 numbers are spaced 1,111 rows apart, and Strip 3 numbers are spaced 111 rows apart. Each strip has a range that is incrementally smaller so that all SBOEID numbers can fit within the same range.

In Yates County, Strip 6 has the widest range, 14,454 numbers. Strip 5 has 14,444 numbers; Strip 4 has 14,344 numbers; and Strip 3 has 13,344. These ranges drop by ten, then a hundred, then a thousand (**Figure 5**). Reducing the range for each strip allows the numbers to nest within each other without overlapping. Quarter and three quarter repunits mark the start and end of each strip with the exception of Strips 1 and 2, which have end values based on the number of registrations in the county.

Strip ID	Strip Min	Distance to Minimum	Gap between SBOEID numbers in strip	Distance to Maximum	Strip Max	Strip Range
1	21,704,366	0	0	14455	21,704,366	14,455
2	21,707,144	2778	1111	5833	21,712,988	5,844
3	21,705,199	833	111	278	21,718,543	13,344
4	21,704,449	83	11	28	21,718,793	14,344
5	21,704,374	8	11	3	21,718,818	14,444
6	21,704,367	1	1	0	21,718,821	14,454
	21,704,387		RegDate after 6/1/2007		21,718,602	

Figure 5: Yates County In-Range strip layout

In deck stacking, the only player who knows which cards are dealt is the player who stacked the deck. Other players see no difference between the cards because they can only see the back of each card. In the same way, individual voter records are anonymized by voter ID numbers. The numbers, like the pattern on the back of a card, have no apparent relationship with the record they belong to. The algorithm adds meaning that would not otherwise be present.

The algorithm positions CID and SBOEID numbers at predictable locations within the 'deck' of voter records. The effect is that every IR number is tagged by the algorithm. Imagine it this way: after an SBOEID sort, John Doe's record, CID #23,765, is in the eleventh position in a group of eleven numbers. That group is the third group of eleven after the sixth group of 111, which is the tenth group of 1,111 and the first group of 11,111.

This alternate method of referencing John Doe's CID number is effectively a third ID number. It could be written as a ten-digit number, where every two digits represents a position, from 01-99 within each of the five strips. The example above would look like this: 0110060311. This algorithmic ID (AID) does not appear in any field in the voter roll database, making it inaccessible for normal use.

A device like this is unnecessary unless access to tagged records must be clandestine. Otherwise, without translating the ID numbers, a system could be designed to map certain ID numbers to any attribute. For instance, the first five numbers in every group of ten could be reserved for people who are vegetarian. However, if that attribute were important to the NYSBOE, it would merit its own field, 'diet'. Instead, there is a well-hidden algorithm creating what is effectively a third ID number.

NYCA's analysis of the OOR partitions is ongoing but what it has discovered to date is enough to link the algorithm with known suspicious records.

Out of range

The OOR partitions contain CID and SBOEID numbers for all of New York's 62 counties. The two partitions are named 'Out of Range Low' (OOR Low) and 'Out of Range High' (OOR High). OOR Low contains 2,436 assigned numbers. OOR High contains 8,940,618 assigned numbers. Because of the relatively small sample of assigned numbers in the OOR Low partition, the analysis presented here is based on data from the OOR High partition.

SBOEID and CID numbers in the OOR partition do not use The Spiral algorithm. OOR numbers are, however, controlled by a different, yet unsolved algorithm. At first, the OOR algorithm, nicknamed 'Tartan' appears designed to randomize numbers. On closer examination, non-random structure is evident. One of these structures, found in Nassau County, led to the discovery that only two values are needed to accurately predict whether a record is purged or not: partition membership and RegDate. Any record with an OOR SBOEID number and a RegDate earlier than 6/1/2007 is almost certainly purged.

The SBOEID number should be unrelated to purged status because all records can have either status. It makes sense that RegDate values would be correlated with purge status because the older they are, the more opportunities there are to purge the record. If that is the reason, then IR and OOR numbers that meet the same RegDate <6/1/2007 criteria should be purged at about the same rate. They are not. There are 11,135,627 records with IR SBOEID numbers and a RegDate <06/01/2007. Of those, 42.45% are purged. There are 710,196 records with an OOR SBOEID number and a RegDate <06/01/2007. Of those, 98.30% are purged (**Table 7**). Records that have a later RegDate are purged in more similar proportions (IR=32.02% vs. OOR=20.17%).

	IR	RD <6/1/2007	Pct	RD >5/31/2007	Pct	OOR	RD <6/1/2007	Pct	RD >5/31/2007	Pct
Purged	4,946,724	4,726,881	42.45%	219,843	32.02%	2,358,943	698,140	98.30%	1,660,803	20.17%
Not purged	6,875,457	6,408,746	57.55%	466,711	67.98%	6,584,111	12,056	1.70%	6,572,051	79.83%
Total	11,822,181	11,135,627		686,554		8,943,054	710,196		8,232,854	

Table 7: Comparison IR vs OOR SBOEID numbers against status and RegDate

A scatterplot of OOR SBOEID and CID numbers can identify purged records at a higher level of accuracy than the previous method. It is, however, more time-consuming. In the scatterplot shown in **Figure 6**, CID numbers are on the X-axis and SBOEID numbers are on the Y-axis. Some records are oriented vertically in the chart (columns); others have a horizontal orientation (slabs). To determine purged status, orientation of numbers in the chart is all that is needed.

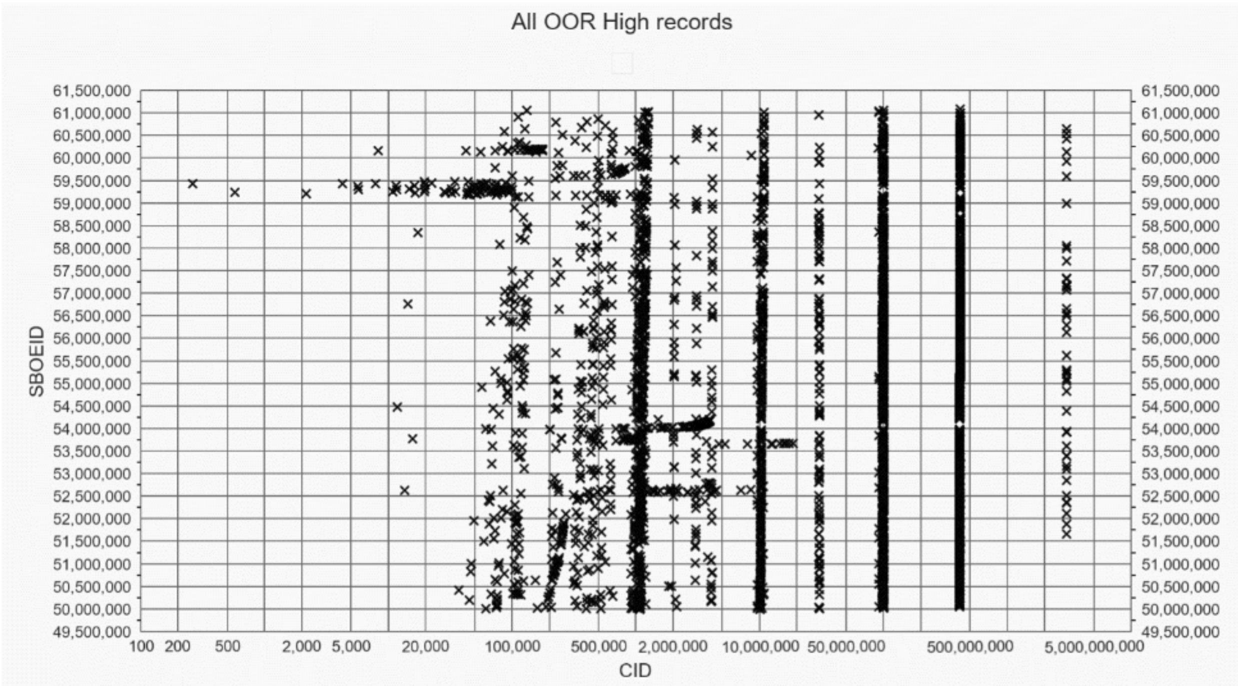

Figure 6: All OOR High partition records

A chart of all OOR partition 'Active' status records illustrates why. In **Figure 7**, purged records have been removed. The effect is that no slabs remain. Every OOR High partition record with active status is in a column.

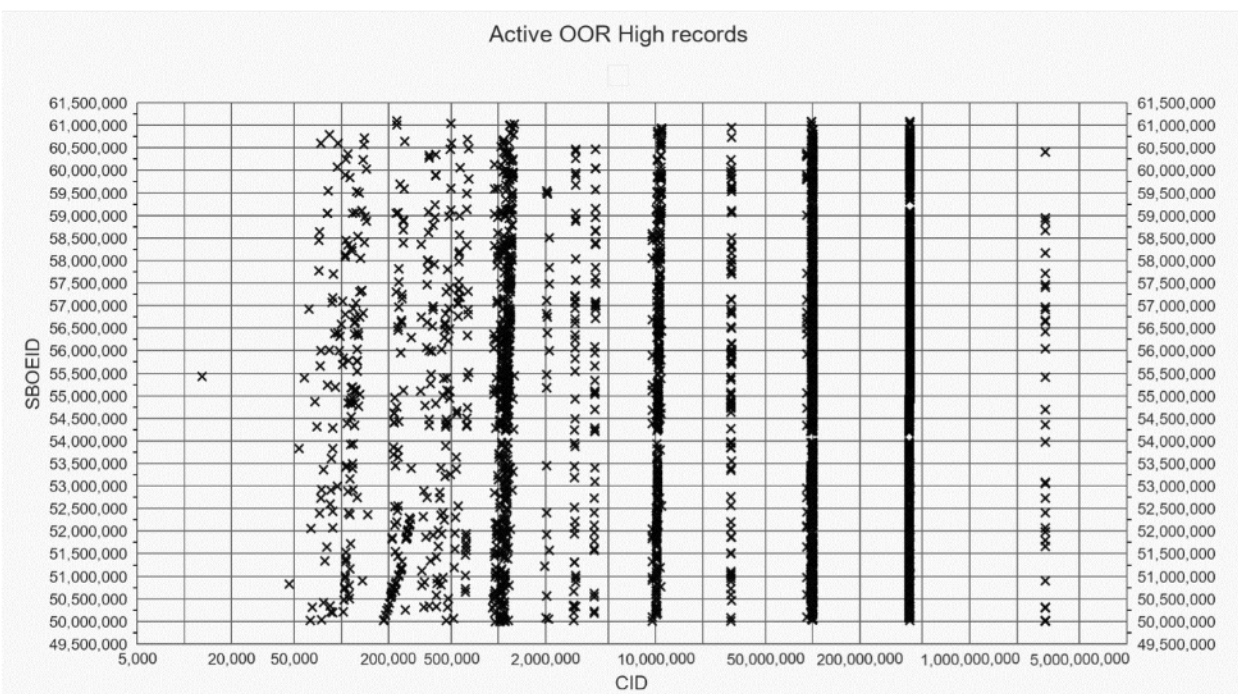

Figure 7: All active records in OOR High partition

This does not mean that all purged records are slabs (**Figure 8**). What it shows is that there are no slab-oriented active records. That suggests that all slab region records were given purged status at the moment of their creation.

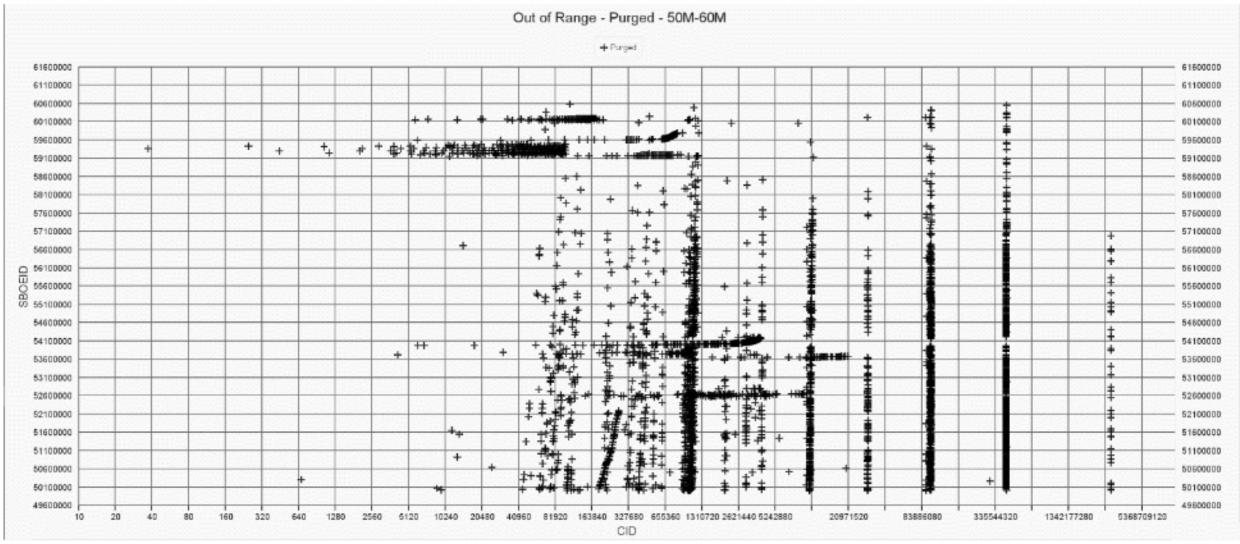

Figure 8: All OOR High purged records show columns and slabs

A close up of some OOR slab numbers from Nassau County reveals an algorithmically produced structure. There are 176,090 records in this group. Approximately 145 of the records are not part of the pattern but occupy nearby space. They are the only active records in this group, and most are unambiguously part of fragmentary column formations. They are, however, too few to be visible in this image. The remaining 175,945 records are purged. Because active records can be distinguished from purged records by formation type, predicting purged status based on formation type is more accurate than predicting status based on RegDate and partition membership.

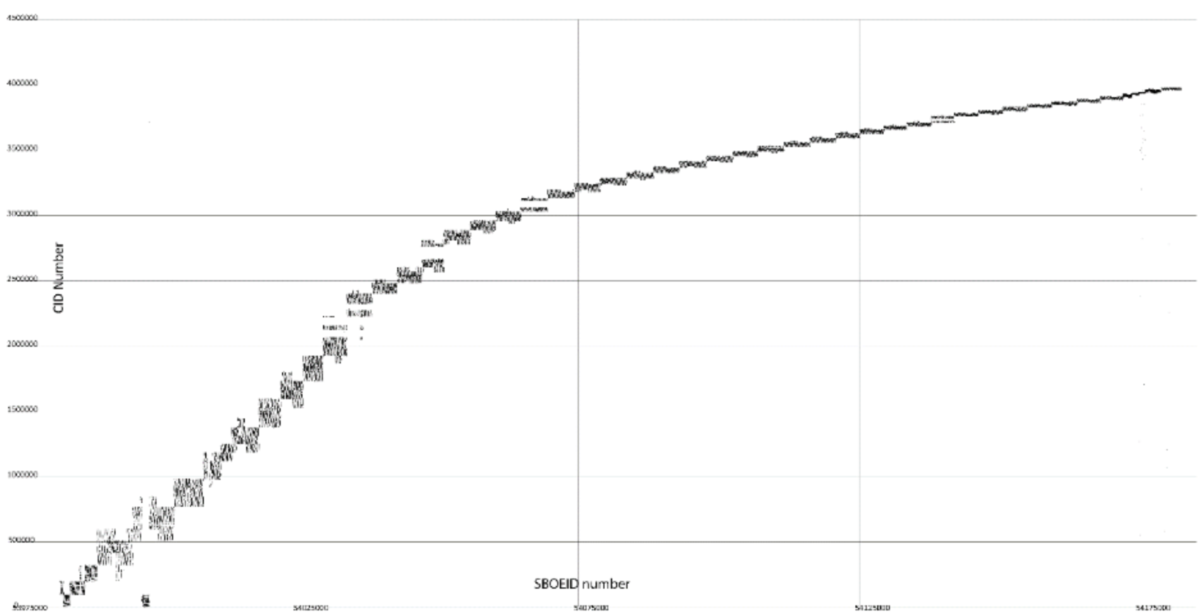

Figure 9: Close-up slab partition, Nassau County reveals graphic structure in number assignments

Another characteristic of slab formation is that a high percentage of records in these regions have been identified as clones (multiple unique SBOEID numbers attached to the same voter). In Nassau County, there are 69,587 clones with a RegDate earlier than 6/1/2007. There is a nearly equal number (*n*=62,971) after that date, but they are part of a much larger group, making them proportionately less numerous than their earlier counterparts (10.75% vs. 28.13%) (**Table 8**).

	Total	<6/1/2007		>5/31/2007	
Clones	132,558	69,587	28.13%	62,971	10.75%
Active	451,006	1,458	0.59%	449,548	76.73%
Inactive	26,345	254	0.10%	26,091	4.45%
Purged	355,902	245,666	**99.31%**	110,236	18.82%
Total	833,253	247,378		585,875	

Table 8: Nassau OOR range clones, before and after 6/1/2007

Conclusion

The Spiral algorithm found in New York's NYSBOE voter roll database is well hidden behind multiple layers of obfuscation.

County ranges are sandwiched between two OOR partitions, where registrations for all counties are mixed. This creates noise in county data that obscures the presence of The Spiral algorithm. Within the IR partitions, counties are not assigned number ranges that follow the order created by county codes assigned by the NYSBOE. This disrupts the logical flow of numbers, making it more difficult to interpret the county ranges if discovered.

The Spiral uses CID numbers to establish the rank order of SBOEID numbers. The result is that a CID number sort does not reveal The Spiral in CID numbers but in SBOEID numbers, and then only if the county does not use alphanumeric CID numbers. An SBOEID sort reveals The Spiral algorithm in CID numbers, regardless of whether alphanumeric values are used.

The Caesar cipher needlessly complicates number sequences by disrupting the relationship between CID number and registration date. The way CID numbers are bound to certain SBOEID numbers further degrades any continuity that might otherwise exist between the numbers. The addition of alternating sequences at the end of Caesar cipher-manipulated number strips makes it more difficult to discern that number sequences have been enciphered.

Stacking the deck by interlacing different number strips buries algorithmic manipulation even deeper. Interlaced cipher strips dramatically increase the overall complexity of the pattern created by The Spiral algorithm, making it less recognizable and less likely to be discovered or understood. A CID sort reveals The Spiral in SBOEID numbers, but only if a calculation is first performed on those numbers.

Normal usage of the voter roll database is unlikely to reveal the algorithm. This is because 'normal' usage does not require the filtering of OOR numbers (likely unknown to any user of the system), downloading a large enough series of consecutively numbered CID or SBOEID records to reveal the algorithm, or filters centered on ID numbers rather than voters' individual personal information, specific geographic locales, or voting districts. For these reasons, it is unlikely that The Spiral algorithm would be discovered by any normal user of the NYSBOE voter roll database.

The fact that no active registrations appear in slab regions is peculiar because purged records are supposedly derived from records that were once active. This presents the possibility that either every person who was assigned numbers in slab regions has since been purged, which is statistically unlikely, or that those regions are reserved for purged records. If so, how did the BOE know that the records were ineligible the very moment numbers were assigned, and why were numbers assigned if they were ineligible?

Nassau's slab section contains 176,090 records. Of those records, more than 99.90% are purged. Among that group, 48,181 (27.38%) have been identified as likely cloned records. Similarly, large proportions of purged clone records have been found in slab sections from other counties. The coincidence of finding these two categories of records together presents a third possibility: the slab regions store records intended for fraudulent usage.

The Spiral algorithm is present in voter roll records obtained by NYCA from the NYSBOE on 21 October 2021. It is also present in the other three versions of the database supplied on 21 May 2022, 26 October 2022, and 21 December 2022. An independent researcher in North Carolina obtained her own copy of NYSBOE voter rolls and checked two counties, Schenectady, and Yates, and found the algorithm in both. The question isn't "is it there?" but "why is it there?".

The algorithm does two things: it restructures SBOEID numbers by binding them to specific CID numbers and hides its presence in the rolls. It has no obvious legitimate utility. It does not make the database easier to use, improve performance, or protect private data. The work involved in creating the algorithm is not trivial. It took experience, time, and ability. This implies that the algorithm has value at least equal to its cost to one or more stakeholders.

There is one problem that the algorithms can solve. It can clandestinely track illegitimate registrations. This solution is necessary for any parties who wish to commit election fraud by casting fake ballots. To prevent an automatic recount or nullification of an election, fraudulent ballots must somehow be reconciled with the number of people who voted in an election. One way to do that is to create fraudulent registrations, then mark those records as having voted in numbers equal to the number of fraudulent ballots. The problem with this method is how to hide fraudulent registrations without losing access to them.

To tag an SBOEID or CID number is simple: add a special set of numbers to signify a fraudulent record. For instance, the NYSBOE uses the last eight digits out of eighteen numbers to identify voters. Nine numbers could be used instead, where every record with a nine-digit ID is fraudulent. However, that would be too easy to find. What can be seen in the voter rolls is subtler, better hidden, and just as effective. Instead of altering numbers, it assigns certain SBOEID numbers to certain CID numbers. This allows the original ID numbers to remain unaltered at the same time they are tagged by associating them with each other.

However, there is no direct relationship between irregular records and algorithm-adjusted AID numbers. This raises the possibility that the algorithms are designed to be one half of an encryption handshake. If so, the algorithm is the key used by an external piece of software that allows access to the records of interest. This only applies to the IR numbers because that is where The Spiral is found. It gives researchers an idea why someone might have done what can be seen in the rolls but does not address what is seen in the OOR partition. OOR numbers are linked to suspicious records in a predictable way.

NYCA's research has uncovered enough anomalies within New York's voter rolls to warrant further investigation. Perhaps that research will find a legitimate purpose for the algorithm. If so, it would be necessary to find an alternate explanation for how the hundreds of thousands of illegal registrations could be used.

References

Atrews, RA 2020, 'Cyberwarfare threats, security, attacks, and impact', *Journal of Information Warfare,* vol. 19, no. 4, pp. 17-28.

NYCA 2022. 'New York's 2020 General Election: A Study in Deficits', ed. M Hornik, AuditNY.com.

Beber, B & Scasso, A 2012, 'What the numbers say: A digit-based test for election fraud', *Political Analysis,* vol. 20, pp. 211-34.

Becker, RA, Volinsky, C & Wilks, AR 2010, 'Fraud detection in telecommunications: History and lessons learned', *Technometrics,* vol. 52, pp. 20-33.

Bolton, RJ & Hand, DJ 2002, 'Statistical fraud detection: A review', *Statistical Science,* vol. 17, pp. 235-49.

Clark, TL 1986, 'Cheating terms in cards and dice', *American Speech,* vol. 61, pp. 3-32.

Fifield, J. 2022. 'RE: Voter ID Assignment Question'. *Private communication, NYCA*.

Francis, RL 1988, 'Mathematical haystacks: Another look at repunit numbers', *The College Mathematics Journal,* vol. 19, pp. 240-6.

Hand, DJ 2010, 'Fraud detection in telecommunications and banking: Discussion of Becker, Volinsky, and Wilks (2010) and Sudjianto *et al.* (2010)', *Technometrics,* vol. 52, pp. 34-8.

Kaur, H & Rani, J 2016 'A survey on different techniques of steganography', *MATEC Web of Conferences,* vol. 57, p. 02003.

Luciano, D & Pritchett, G 1987, 'Cryptology: From Caesar ciphers to public-key cryptosystems', *The College Mathematics Journal,* vol. 18, pp. 2-17.

National Voter Registration Act 1993, Public Law 103-31, 103rd Congress.

NYSBOE 2021, 'New York State Board of Elections voter rolls', version current on 12 October 2021.

NY 2021 Election Law §1-6219 2021, viewed August 2021, <https://www.elections.ny.gov/ElectionLaw.html>.

Sikdar, K 1992, 'Generalized t-ary trees and their path lengths with applications', *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002),* vol. 54, pp. 443-59.