Assessment of the Relationship Between Collision Rate and Sample Size Using a Large US Mortality Dataset

114312

Huynh S[1], Liu T[2], Leshin J[2], Haskell T[1]

[1]Cerner Enviza (Previously Kantar Health), Malvern, PA, USA; [2]Datavant, San Francisco, CA, USA

Introduction

- Storage of personally identifiable information (PII) such as full name, social security number (SSN), addresses, and date of birth can be easily used to compromise an individual's privacy, necessitating the need for methods to safeguard privacy.
- A common de-identification technique for obscuring PII is the use of a cryptographic function known as a hash function on the PII; hash codes are generated from different combination patterns of PII elements.[1]
- Patient PII is typically hashed to preserve privacy; however, PII collisions can introduce false-positive matches in patient analyses when connecting patient records across multiple datasets.

Objectives

- To estimate expected collision rates in a large United States (US) mortality dataset and specifically examine the relationship between collision rate and sample size.
- We propose the hypothesis that expected collision rate scales linearly with sample size.

Methods

Data Source

- We used data from a large US mortality dataset based on government data sources with over 100 million unique individuals.
- Datavant's Death Index is a mortality database with information on deceased individuals in the US and Canada. It contains information on individual deaths and is updated on a weekly basis.
- The data coverage spans from 1937 to the present with varying proportions of full mortality coverage through the years.
- The data includes individual-level information such as age at death (**Table 1**).
- In this analysis, we look at Token 1 and Token 2.
- A collision between unique individuals in the dataset is defined by a match of Datavant Tokens, which are defined by combinations of PII elements. Here, we show results for Tokens 1 and 2, which are defined by:
  - **Token 1:** Last Name + First Initial of First Name + Gender + Date of Birth
  - **Token 2:** Last Name (soundex) + First Name (soundex) + Gender + Date of Birth

Statistical Analysis

- A series of random sample populations were drawn from the dataset. At each sample size, we compute:
  - The number of unique Tokens (K) belonging to these unique individuals (**N**).
  - The number of collisions (C), defined as N-KWe compute our results as a function of the estimated total number of possible Token values in our population (D), which can be derived from observed values of C and N.
  - The expected collision rate (**C/N**).

Table 1. Distribution of Patient Age at Death in the Mortality Data

Age Range (Years)	Percentage
< 18	0.57%
19-40	5.22%
41-55	10.03%
56-65	14.02%
66-75	22.25%
76-85	27.16%
86+	20.75%

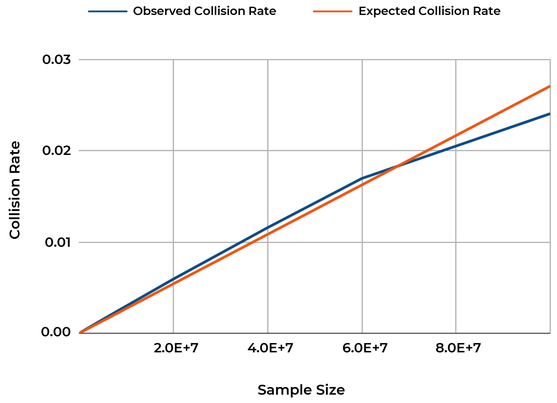
Results

- The sample size threshold for at least one expected collision was found to be between N = 40,000 and N = 70,000 (**Table 2**; **Figure 1**).
- We observed that as sample size increased, the number of collisions and collision rate correspondingly increased (**Figure 2**).

Table 2. Collision Rate by Sample Size (Using Token 2)

Sample Size (N)	Unique Token 2 (K)	Collisions (C)	Collision Rate (C/N)	Fitted Collisions N(N-1)/2D	Fitted Collision Rate (N-1)/2D
10,000	10,000	0	0	0	0.000003
20,000	20,000	0	0	0	0.000005
30,000	30,000	0	0	0	0.000008
40,000	40,000	0	0	0	0.000011
70,000	70,000	0	0	1	0.000019
100,000	99,999	1	0.00001	3	0.000027
200,000	199,995	5	0.000025	11	0.000054
300,000	299,982	18	0.00006	24	0.000081
500,000	499,923	77	0.000154	68	0.000136
800,000	799,810	190	0.000238	173	0.000217
1,000,000	999,693	307	0.000307	271	0.000271
2,000,000	1,998,785	1,215	0.000608	1,084	0.000542
3,000,000	2,997,284	2,716	0.000905	2,440	0.000813
5,000,000	4,992,481	7,519	0.001504	6,777	0.001355
9,000,000	8,975,735	24,265	0.002696	21,957	0.00244
10,000,000	9,970,029	29,971	0.002997	27,108	0.002711
20,000,000	19,881,212	118,788	0.005939	108,430	0.005422
40,000,000	39,536,552	463,448	0.011586	433,720	0.010843
60,000,000	58,981,714	1,018,286	0.016971	975,870	0.016265
100,000,000	97,289,249	2,408,121	0.02408121	2,710,751	0.027108

Figure 1. Collision Rate by Sample Size (Using Token 2)

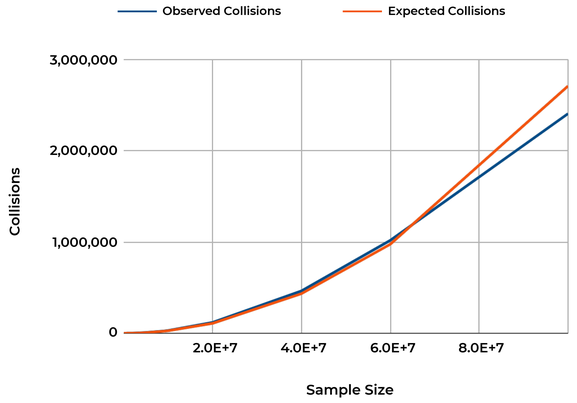


- When Token 1 and Token 2 were used together, the count of estimated distinct PII combinations in the population (D) were higher compared with using Token 2 individually (5.65 billion vs. 1.845 billion, using 100% of the dataset sample), and the collision rate was lower (**Table 3**; **Figure 2**). This is to be expected, as Token 1 and Token 2 together contain more information than Token 2 alone.

Table 3. Collision Rate by Sample Size (Using Combined Token 1s and Token 2)

Sample Size (N)	Unique Token 2 (K)	Collisions (C)	Collision Rate (C/N)	Fitted Collisions N(N-1)/2D	Fitted Collision Rate (N-1)/2D
10,000	10,000	0	0	0.01	0.000001
20,000	19,999	1	0.00005	0.04	0.000002
30,000	29,999	1	0.000033	0.08	0.000003
40,000	39,999	1	0.000025	0.14	0.000004
70,000	69,999	1	0.000014	0.43	0.000006
100,000	99,997	3	0.00003	0.88	0.000009
200,000	199,994	6	0.00003	3.54	0.000018
300,000	299,993	7	0.000023	7.96	0.000027
500,000	499,979	21	0.000042	22.12	0.000044
800,000	799,931	69	0.000086	56.64	0.000071
1,000,000	999,894	106	0.000106	88.5	0.000088
2,000,000	1,999,605	395	0.000198	353.99	0.000177
3,000,000	2,999,165	835	0.000278	796.47	0.000265
5,000,000	4,997,604	2396	0.000479	2212.42	0.000442
9,000,000	8,992,208	7792	0.000866	7168.23	0.000796
10,000,000	9,990,383	9617	0.000962	8849.67	0.000885
20,000,000	19,962,136	37,864	0.001893	35,398.68	0.00177
40,000,000	39,850,765	149,235	0.003731	141,594.72	0.00354
60,000,000	59,670,338	329,662	0.005494	318,588.12	0.00531
100,000,000	99,113,618	886,382	0.00886382	884,967	0.00885

Figure 2 . Collision Count by Sample Size (Using Token 2)



Limitations

Number of Unique Matches and the Expected Number of Pairings

- This methodology does not account for the number of unique matches or the expected number of pairings made.
- For example, if 3 unique individuals A, B, and C happen to share the same Token, there are 3 matching pairs (A,B), (A,C), and (B,C) which are all false-positive matches while the number of collision C = N – K = 3 – 1 = 2.

Distribution of PII Fields

- For the derivation of the proposed formula, we must assume that the probability of a certain Token value being generated from an individual's PII fields is uniformly distributed across the underlying population.
- However, unlike the case of Birthday/Hashing collision problems,[2,3] this assumption is not necessarily true for our Token collision problem since the distribution of PII such as name and date of birth are not uniform. Specifically, for example, the probability of another person in the US sharing the same Token 2 with John Smith vs. a less common name will be different.

Conclusions

- The hypothesized relationship between collision rate and sample size was validated using a large US mortality dataset.
- The results indicate that collision rate scales about linearly with sample size.
- This validation helps to further inform how the false-positive rate of Token-based matching algorithms may change with sample size.

References

1. Chen X, Fann YC, McAuliffe M, Vismer D & Yang R. (2017). Checking Questionable Entry of Personally Identifiable Information Encrypted by One-Way Hash Transformation. *JMIR medical informatics*, 5(1), e5054.
2. Stein C, Drysdale RL & Bogart KP. (2011). Discrete mathematics for computer scientists. Boston, Mass: Pearson/Addison-Wesley.
3. Fadnavis S. (2011). A generalization of the birthday problem and the chromatic polynomial. arXiv preprint arXiv:1105.0698.

Presented at ISPOR Europe 2021, November 30-December 3

Email: Stephanie.Huynh@cernerenviza.com