WHITEPAPER

The Trouble with Names/ Dates of Birth Combinations as Identifiers

Joseph R. Barr / Stephen Coggeshall / Wenzhong Zhao

April 2011

Table of Contents

Abstract

Direct marketing and many other consumer analytics frequently rely on name/dates-of-birth (DOB) combinations for personal identification. We will demonstrate that empirically, names/DOB combinations are woefully inadequate as a precise tool for personal identifiers; this is true for common, especially for cyclical and trendy names appended by a DOB. Although we will explore the ball-in-cell paradigm as a plausible methodology, this paper is focused on the empirics of names, first, last and full-names, and, names-DOB distribution in the U.S. and its relation to the dimension of time. Although various sources of data are in existence, this particular study is based on the proprietary ID Analytics identity resolution database, one consisting of over 300 million high-quality persons' records.

Introduction

With a substantial return on investment, direct marketing of goods and services is a growth industry with an estimated 2010 expenditure (by retailers, etc.) of just under $200 billion annually. For practical reasons direct marketers like InfoUSA, Acxiom, Choice Point and Equifax heavily rely on the name-DOB combinations for personal identification. Embedded in this is the assumption that name-DOB combinations will, in all but rare cases, uniquely identify the intended person. Personal identification is related to the celebrated birthday problem; simply stated, it says that in a group of 23 people the chance that two will share a birthday, month-day only, is approximately 50%. The name-DOB identification problem is related to the chance for date *collision* is concerned with the probability that in a group of *M* people, two share a birthday (day, month and year). In this mostly empirical study we will debunk the seemingly reasonable notion that it is possible to uniquely identify a person (in the U.S.) by the first name, more precisely, first-name-root (FNR), last name and date-of-birth (DOB) combinations – that identification fails for practically all common American names. Furthermore, we will show that although the aggregate of all names (first-last combination) is quite large, the name space isn't uniformly distributed – not even nearly so – therefore no reasonable approximation method could possibly work; furthermore, the name space distribution is highly skewed: a non-trivial group of first names (represented by FNR) follow a cyclical patterns: it's as uncommon to find a young Wilbur or Harold as it is to find an old Brittany or Kyle. For example, according to the Social Security Administration (SSA), although not common for men, Kelly is a shared male/female name; the girl name Kelly enjoyed great popularity between 1970 and 1980 reaching its zenith (ranked #10) in 1977 before it began to wane to become a relatively rare name.

1. Mathematical Background

a. Balls in Bins

The balls-in-cells paradigm, first studied by Richard von Mises, is concerned with the statistical properties of occupancy where M indistinguishable balls are randomly cast into N indistinguishable cells. The salient assumption is that a ball is equally likely to be in any given cell with equal probability $1/M$. A well-suited analogy to our version of the birthday problem, where the year is included, is one where the set B of M balls is partitioned into 'color' classes $C_1,\ldots,C_k$ of balls is colored with the *k* different colors: color 1 through color *k*. The setup implies that $\left|C_j\right| = m_j$ and $\sum_{j=1}^{k} m_j = M$. The analogy with the name-DOB identification problem is clear: the balls map to the persons, the colors to the names and cells to the dates. Calculations are greatly aided by the method of indicator. Let $X_{c,j}$ = the number of *j*-colored balls in some cell *c*. We seek to calculate $\Pr\left(X_{c,j} \geq 2\right)$. Consider the balls colored *j* (say red) $b_{j,1},\ldots,b_{j,m_j}$ and define the independent indicator $\mathrm{I}_{c,b_{j,l}} = 1$ if the *l*th *red* ball, i.e., in the color class C_j , is in cell *c*, and $\mathrm{I}_{c,b_{j,l}} = 0$, otherwise.

Clearly $X_{c,j} = \sum_{i=1}^{m_j} \mathrm{I}_{c,b_{j,i}}$ and so the expected number of *red* balls in cell *c* is

$EX_{c,j} = \sum_{i=1}^{m_j} E\mathrm{I}_{c,b_{j,i}} = \sum_{i=1}^{m_j} \Pr\left(\mathrm{I}_{c,b_{j,i}} = 1\right) = \frac{m_j}{N}$. Collision occurs in cell *c* if $X_{c,j} \geq 2$, if two red balls are in cell *c*. Now if m_j strictly exceeds *N*, then the pigeonhole principle guarantees collisions, that is two or more red balls in cell *c*. For color classes of cardinality *m* where $m < N$ the probability that a cell *c* will contain two identically colored balls is

$$\frac{1}{N^m}\sum_{j=2}^{N}\binom{m}{i}(N-1)^{m-j} = \left(\frac{N-1}{N}\right)^m \sum_{j=2}^{m}\binom{m}{i}(N-1)^{-j}$$

To demonstrate this idea consider *N*=21,915 (number of days in 60 years, between 1931 and 1991), *m*=10,009, in fact this is not at all a fictitious number, ID Analytics data identify 10,009 distinct persons all named Kenneth Jones. The probability that two Kenneth Jones share any particular birthday is 0.148. If we assume DOB of Kenneth Jones' to be jointly independent, the probability that at least two Kenneth Jones share a birthday: the probability that in any arbitrarily chosen 30 of the 21,915 dates is approximately 99.2%. The calculations are based on the probability that at least two Kenneth Jones were born on same date is equal to 1 less the probability that no two Kenneth Jones were born on same date, specifically

$$\Pr(\bigcup_k C(k)) = 1 - \Pr\bigcap \overline{C}(k) = 1 - \prod_k \Pr\overline{C}(k) = 1 - \prod(1 - \Pr(C(k))) .$$

Where $C(k)$, $\overline{C}(k)$ stand for collision/no collision on date *k*.

b. Theoretical Probability of Collisions

The theoretical probability of collision will be used throughout. In this section we calculate the value of that probability. Assume a particular name class consists of *m* distinct persons, and, let *N* be the number of days (or dates) in the time horizon in question, and, let $f(j) =$ the ratio of the average number of births on date *j* divided by population size (approximately 308 million). One can think of

$f(j)$ as the probability to be born on date j. With respect to earlier definitions of $C(k)$ and $\overline{C}(k)$ and $D(j), \overline{D}(j)$ the event "was born on date j", one can easily follow the sequence of formulas

$$\Pr \overline{C}(k) = \left(\frac{N-1}{N}\right)^m$$

$$\Pr(\bigcup_k C(k)) = 1 - \left(\frac{N-1}{N}\right)^m$$

$$\Pr(\overline{D}(j)) = 1 - f(j)$$

$$\Pr \bigcap_{j=1}^{N} \overline{D}(j) = \prod_{j=1}^{N} (1 - f(j))^{m \cdot f(j)}$$

Since a collision is equivalent to "two or more people having the same name were born on the same date" we have that the probability of a collision (on some date k) equals to

$$\Pr \bigcup_k C(k) = 1 - \prod_{j=1}^{N} (1 - f(j))^{m \cdot f(j)} \text{ QED.}$$

2. The Data

The name-DOB analysis is performed against ID Analytics' 308,707,122 records in the Identity Resolution database. The data covers 111 years of dates of birth, commencing at the first of the year 1900 through the end of the year 2010, a total of 40,542 distinct dates. Following standard industry practices, the data regularly undergoes a rigorous QA regime which includes sophisticated de-duping to ensure no individual does occupy anything other than a single record in our database. Based on advanced machine learning technology, not only can ID Analytics identify a Larry Harper as a Lawrence Harper but also as one that goes by the occasional nicknamed "Bud" Harper, a name he was known by and extensively used by the media, a remnant from his college football days.

3. First Names Distribution

a. First Names Aggregates

Every generation has stories to tell about the number of 'Johnny's' in their classes. Anecdotally the chance for multiple kids having identical first names seems rather high. The data confirms this anecdotal notion as a fact. As suspected, the first name space is quite large, consisting of 3,540,824 mostly rare distinct names. At the two extremes are the common names like John, James, Mary, Jane and Robert and the other extreme are the uncommon names, those shared by fewer than 1,000. Interestingly, top-10 names (Table 1 below) account for over 18 percent and top-50 account for nearly 50% of the population.

Table 1. Top-10 first names.

Rank	First Name	Frequency
1	JOHN	7,556,152
2	MARY	7,474,295
3	JAMES	5,714,116
4	ROBERT	5,497,484
5	MICHAEL	4,942,065
6	CHRISTOPHER	4,747,669
7	WILLIAM	4,665,950
8	JOSEPH	4,619,701
9	ELIZABETH	4,270,062
10	RICHARD	4,109,367
	TOTAL	**53,596,861**

Figure 1 illustrates a dramatic first names utilization phenomenon; we can see that a tiny fraction of the names occupy well over 90 percent of the population. In fact, this is so dramatic, the graph is presented for no more than a wow factor.

Figure 2 depicts the phenomenon somewhat definitely.

Figure 1. First names density.

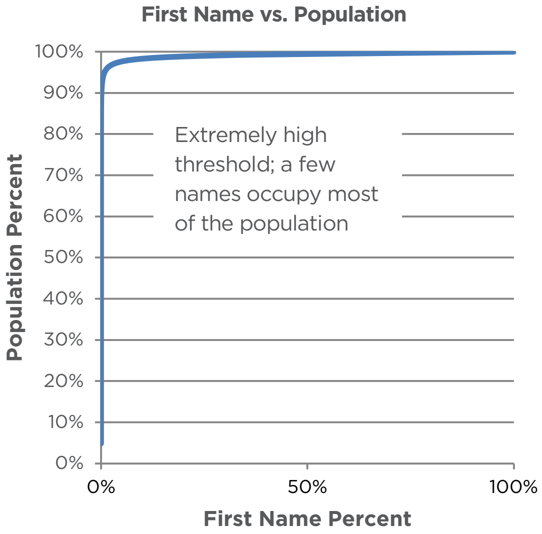
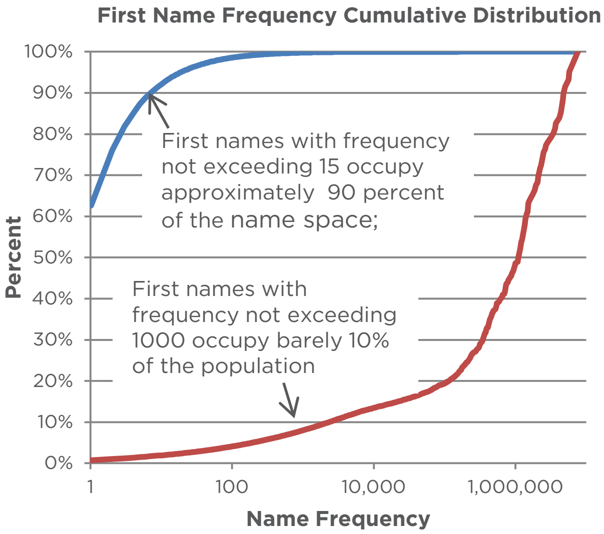


Figure 2. First name distribution.



b. Trendy Names

This section will illustrate name trendiness by examining the year of birth of two common names, Kelly Johnson and Jason Smith. One can't help noticing the narrow band of the popularity; both Kelly and Jason were popular in the 1970s peaking at about the same time period. Figure 3 and 4 makes our point.

Figure 3. "Kelly Johnson"

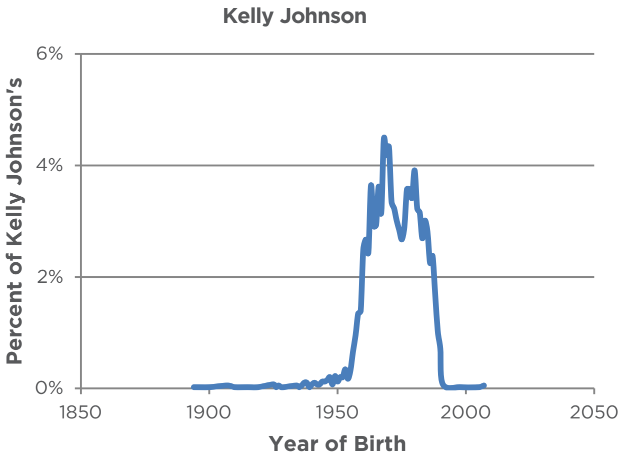
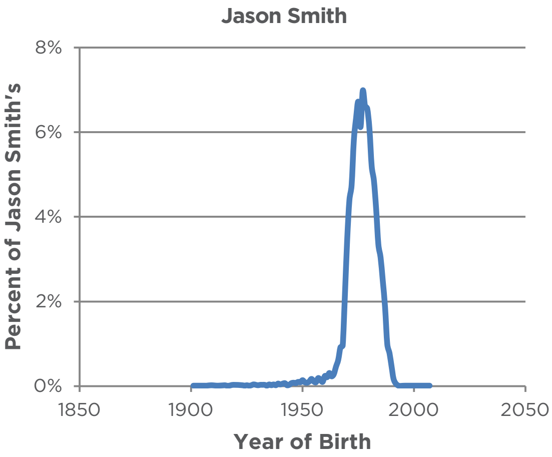


Figure 4. "Jason Smith"



4. Last Names Distribution

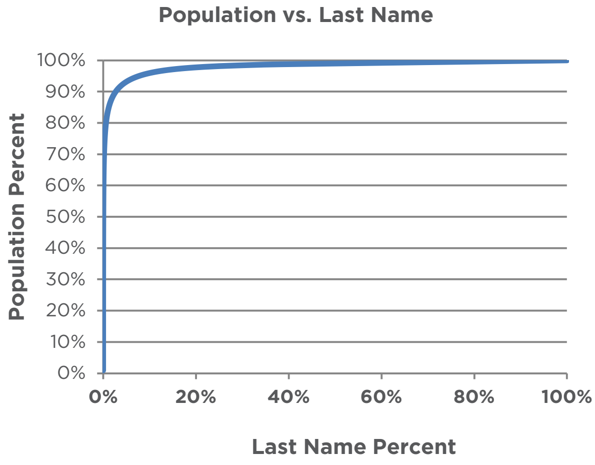
Although last names space is vast consisting of 6,312,209 distinct last names, only a small fraction of those account for the majority of the population. Specifically, 21,738 distinct last names, or 0.34% of the space of last name account for approximately 75% of total. Not surprisingly although last names exhibit a strong threshold property, it is not quite as dramatic as first names. Displayed in Table 2 is a last names top-10 list occupying 16,043,057 persons or about 5 percent of the population; the top-50 last names occupy 40,345,230 persons or about 13 percent of total.

Table 2. Top-10 last names.

Rank	First Name	Frequency
1	SMITH	2,848,936
2	JOHNSON	2,251,054
3	WILLIAMS	1,914,430
4	JONES	1,724,034
5	BROWN	1,671,484
6	DAVIS	1,260,597
7	MILLER	1,233,005
8	GARCIA	1,093,491
9	RODRIGUEZ	1,053,622
10	MARTINEZ	992,404
	TOTAL	**16,043,057**

The dramatic nature of last names threshold phenomenon is displayed in Figure 5. One notices that top 1% of last names account for over 90% of the population.

Figure 5. Last name distribution as percent of the population.



5. Full Name Distribution

Going down the food chain, the distribution of full names, first plus last, is further flattened. Top-10 of full names occupies a mere 550,014 or 0.2% of the U.S. population, while top-50 occupies 2,040,991 or approximately 1% and 5,691 full names occupy slightly over 10% of total. Table 3 displays top-10 list of full names.

Table 3. Top-10 full names.

Rank	Last Name	First Name	Frequency
1	SMITH	JOHN	64,794
2	SMITH	JAMES	64,180
3	SMITH	ROBERT	57,004
4	RODRIGUEZ	MARY	55,565
5	GARCIA	MARY	54,605
6	SMITH	WILLIAM	54,246
7	GARCIA	JOSEPH	50,853
8	HERNANDEZ	MARY	50,197
9	MARTINEZ	MARY	49,356
10	RODRIGUEZ	JOSEPH	49,214
		TOTAL	**550,014**

Driving this phenomenon is a vast name space of 75,568,646 full names. In fact, full names with frequency less than 40 consist of approximately 50% of the U.S. population while names having frequencies greater than 500 account for nearly 25% of total. Table 4 displays the bottom 10 names of frequencies 1 through 10.

Table 4. Distribution of infrequent names.

Frequency	Count	Percent	Cum Percent
1	53,814,135	17.43%	17.43%
2	19,564,894	6.34%	23.77%
3	11,003,037	3.56%	27.33%
4	7,576,004	2.45%	29.79%
5	5,822,960	1.89%	31.67%
6	4,772,826	1.55%	33.22%
7	4,034,289	1.31%	34.53%
8	3,526,472	1.14%	35.67%
9	3,126,492	1.01%	36.68%
10	2,819,760	0.91%	37.60%
TOTAL	**116,060,869**		

Figure 6, representing full names against populations, is similar to earlier ones with the notable exception that a higher percentage of full names would be required to account for a lion share of just over 50% of the population. Approximately 20,000 full names, or 0.03% of full names are required for 50% population coverage.

Figure 7 is similar to Figure 2 in depicting full names distribution with respect to the total namespace and U.S. population.

Figure 6. Population vs. full names.

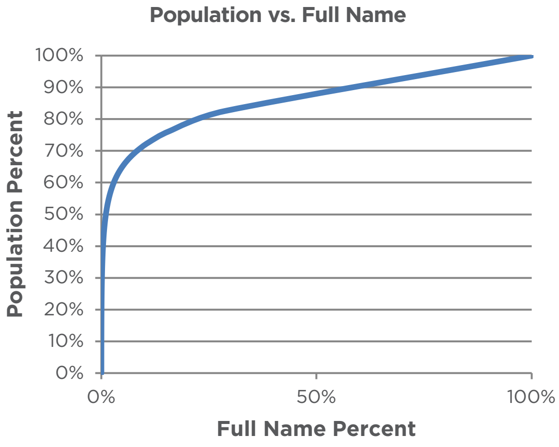
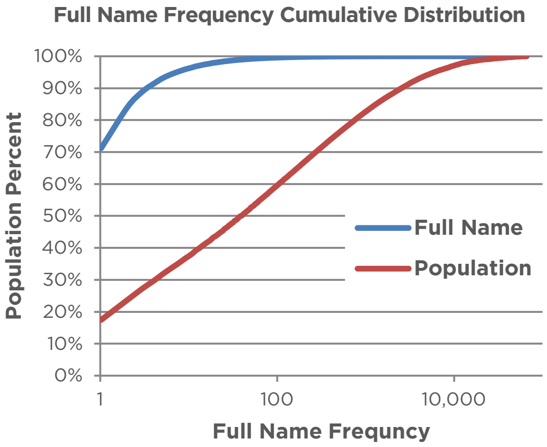


Figure 7. Name frequency against name space and population.



6. Name-DOB Distribution

Not surprisingly, popular names are predisposed to collisions. To illustrate, name-DOB collision within the three popular names John Smith, James Smith and Robert Smith is approximately 79%, thus of the aforementioned names, only 21% could be uniquely identified by name-DOB combinations. Evidently, uncommon names are less likely to collide than common names and unique names (one not shared by two or more persons) are by definition invulnerable to collision.

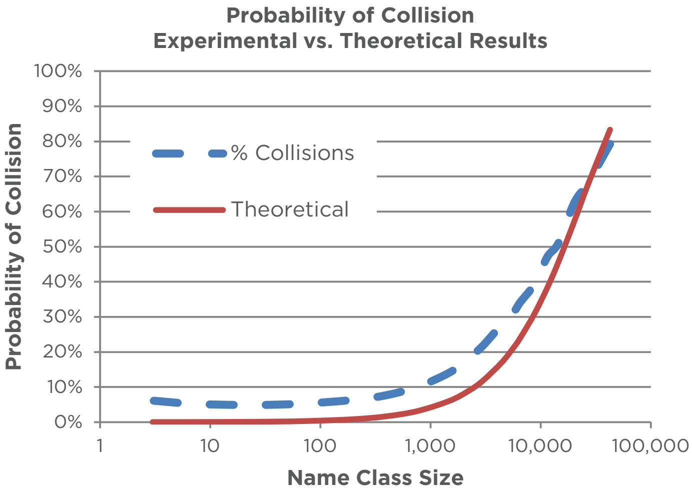
Table 5 displays collisions where names are aggregated by frequencies. Notice that with the exception of those names shared by fewer than 400 persons, the chance for collision is non-negligible.

Table 5. Theoretical vs. actual number of collision with names aggregated.

# Names in Group	# Persons in Names Group	Average Persons Per Name	Percent Colliding	Probability Colliding
3	127,901	42,633	79.27%	83.34%
4	107,799	26,949	68.19%	67.79%
6	120,738	20,123	62.44%	57.09%
8	115,853	14,481	50.98%	45.60%
10	115,295	11,529	47.16%	38.41%
12	105,030	8,752	39.30%	30.78%
15	97,646	6,509	34.12%	23.94%
20	115,358	5,767	31.63%	21.53%
25	111,415	4,456	27.94%	17.08%
49	146,919	2,998	21.82%	11.84%
65	145,888	2,244	18.33%	9.00%
106	159,134	1,501	14.20%	6.12%
215	163,222	759	10.05%	3.14%
437	167,155	382	7.57%	1.59%
1,031	197,856	191	6.28%	0.80%
3,187	245,669	77	5.34%	0.32%
12,489	482,115	38	4.97%	0.16%
47,194	911,859	19	4.90%	7.98E-04
281,976	2,167,062	7	5.21%	2.94E-04
1,164,592	4,430,043	3	6.12%	1.26E-04

Figure 8 is a dramatic depiction of probability of names-DOB collision as a function of the size of the name equivalence class. Also notice how the curves representing expected and actual move in unison. Also notice that the probability exceeds 10% for name classes in the 2,500 range and higher.

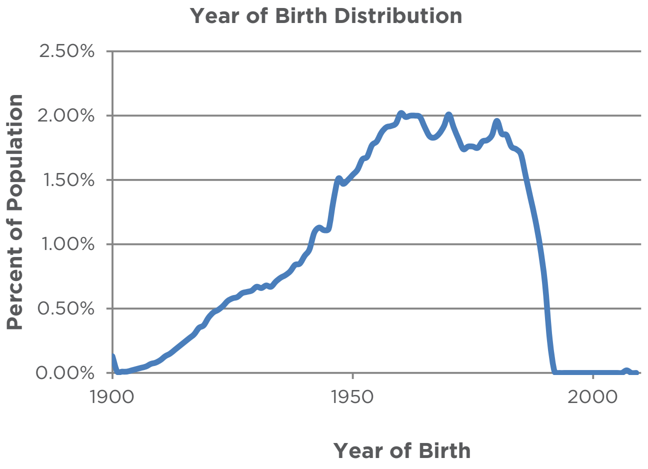
Figure 8. Probability of name-DOB collision as a function of name frequency.



7. The Time Dimensionality

In devising statistical models for name-DOB distribution the time dimensionality cannot be ignored. Figure 9 depicts the age distribution of the credit-worthy population, mostly ages 20 through 90.

Figure 9. Age distribution.



Conclusion

This study, using actual data, has demonstrated the problem of using name-DOB combinations to uniquely identify a person. What is, however, surprising is that despite all that was said, it's still possible to uniquely identify a substantial portion of the population. In fact, almost 92% of the people in our database are uniquely identifiable solely by their names and dates of birth and so one might want to ask the question why bother with the remaining 8.3% who aren't uniquely identifiable (by names and DOB.) Before one can answer the question, one must ask what purpose accurate identification serves, and additionally, what is the cost associated with misidentification. Obviously this question has many answers, but we will only mention a couple of points to amplify the issue

a. According to MINTEL/Comperemedia, $2 billion was spent in 2010 in direct marketing of credit card mailing.

b. Although not very common, during recent decades, state agencies reliance on name-DOB combinations have known to produce unpleasant, sometimes catastrophic outcomes for persons who were misidentified as criminals or other type of law-breakers.

We tend to agree that the issues are complex and the dollars and social costs of making identification "fool proof" are rather prohibitive, yet we believe that improving persons' identification methods and technology will increase ROI, especially in the presence of imperfect and incomplete inputs.

Acknowledgment

The first author would like to thank Mr. Patrick Reemts for his helpful suggestions.

References

1 ID Analytics *Identity Resolution* database

2 The Open Group, *Identity Management*, http://www.opengroup.org/projects/idm/uploads/40/9784/idm_wp.pdf

3 U.S. patents 7,458,508; 7,562,814; 7,686,214; 7,793,835

www.idanalytics.com