CONSEQUENCES OF SOCIAL SECURITY NUMBER RANDOMIZATION

Bert Kestenbaum

Social Security Administration
Bert.M.Kestenbaum@ssa.gov

Introduction

On June 25th of last year, the Social Security Administration (SSA) abandoned the method for assigning social security numbers, or ssns, that had been in place for three-quarters of a century in favor of random assignment. Of the one billion possible nine-digit numbers, about 466 hundred million numbers had been assigned by June 24th, 2011. Since the 100 million numbers that begin with the number „9" and the two million which begin with „000" or „666" and another 9 million with „00" as digits 4 and 5 are still not being used, the balance of 423 million records formed the pool of available numbers from which the computer"s random number generator will draw. As of today, approximately 4 million numbers have been assigned by the new method.

The impetus for the changeover to random assignment came largely from concerns that the regularities inherent in the old assignment method could be exploited for identity theft. For example, in a paper published in the Proceedings of the National Academy of Sciences in 2009, the authors demonstrated that with knowledge of an individual"s date of birth and state of residence, one could, by applying these regularities in ssn assignment to the SSA"s publicly-available Death Master File, often determine a fairly narrow interval in which the individual"s social security number is likely to fall (Acquisti and Gross 2009). At one point the concern that SSA would shortly run out of numbers was being used to argue for a change to randomization. However, our Office of the Chief Actuary calculated, after applying appropriate assumptions about the levels of future births and immigration, that the current method, if applied efficiently, could be used for another 50 years before the collection of one billion possible numbers is exhausted, and, furthermore, that the date of exhaustion under a randomization scenario would be only slightly later.

Social security numbers assigned before June 25, 2011 had the following structure:

- The first three digits, called the area number, indicate the applicant"s State of residence. One or more area numbers are allotted to each State. For example, areas 212 through 220 are allotted to Maryland. Areas 700 through 728 were allotted to the Railroad Retirement Board for ssn assignment up through June 1963 to persons covered under the Railroad Retirement System. Area 586 was allotted for persons not residing in the United States.

- The next two digits, called the group number, have information about when the ssn was assigned. The group number follows this sequence: 01, 03, 05, 07, 09, all the even numbers from 10 to 98, then the even numbers from 02 to 08, and then all the remaining odd numbers, beginning with 11. For a State with, say, an allotment of 5 area numbers, the first 50 thousand numbers assigned have the group number „01", the next 50 thousand have group number „03", and so forth. The Social Security Administration periodically published the "High Group Listing", which showed for each area number what the current group number was. This listing could be used by employers or businesses to determine that a social security number presented by a prospective employee or customer was incorrect or bogus – simply by noting that the group number was later in sequence than the group number on the Listing. From the final publication of the Listing, for issuances through June 24, 2011, one can deduce which areas were fully assigned, partly assigned, or never assigned under the old method.

- The last four digits are called the serial number. Within an area-group, numbers are assigned according to the following sequence: 0001, 0002, 0003, 0004, 2000, 0005, 0006, 0007, 0008, 2001…. This pattern was

chosen to ensure that every 5th serial number assigned from a particular area-group begins with either a „2" or a „7", which was thought to be important to SSA"s statistical sampling system. It is noteworthy that this peculiarity in the sequencing of serial numbers was not known to the aforementioned authors of the paper on identity theft. Also, it is clear that this peculiarity did not exist from the beginning of ssn assignment in 1936; it seems likely that it started in 1972, when SSA discontinued the issuance of ssns by field offices in favor of central issuance from its Baltimore headquarters.

A discussion of the assignment of social security numbers would be incomplete without mention of the Enumeration at Birth (EAB) program. In this process, the parents apply for a social security number for the infant at the same time the birth is registered. The rollout of EAB began in 1987, and a few years later all States were participating. (More recently, SSA put into place an Enumeration at Entry process, so that an application for a social security number is part of the paperwork for an immigrant upon his arrival in the United States.) The incorporation of EAB into the enumeration process exacerbates the identity theft vulnerability: if an identify thief knows a target"s date and State of birth, and he finds a record in the Death Master File with the same date and State and an ssn assigned by EAB, then he knows that the target"s ssn is likely to be very close to that one.

While the changeover to randomization is necessary for the protection of personal identities, there are unfortunate ramifications for demographic research, in terms of lost information. An obvious example is the loss of the geographic information implicit in the first three digits of the ssn. Also, the group number, taken together with the area number, reflects when the ssn was issued, and could be used, for example, together with the numberholders" age to infer that the numberholder is likely foreign-born. Some time ago, Gloria Block and her colleagues "differenced" High Group Listings to ascertain the years in which the social security numbers in each area-group block were issued (Block, Matanoski, and Seltser 1983). I have provided similar, but up-to-date information to researchers who have requested it, based directly on information in SSA"s master file of ssn applications.

In the balance of this paper I treat two areas of research which can be explored because of the characteristics of the old, non-random method of social security number assignment.

Twins

Studies of twins are useful for, among other things, measuring the relative importance of genetics and the environment in disease and other outcomes. Twin research in countries in Europe and Asia that have population registers is quite extensive, but the same cannot be said for the United States. Several years ago, the National Institute of Environmental Health Sciences at the National Institutes for Health initiated a study of the feasibility of creating a national, population-based twin registry in the United States. I haven"t seen anything beyond an interim report, and clearly there was no sustained movement towards creating such a resource.

The most successful twin registry created in the United States remains the National Academy of Sciences-National Research Council twin registry of World War II military veterans, consisting of almost 16 thousand pairs of white male twins born between 1917 and 1927 who served in World War II (Jablon *et al.* 1967). This registry contains information on zygosity – that is, whether the twins are identical or fraternal – solicited by the question, "Have you ever been told that you look like two peas in a pod?"

Several years ago, the fact that the WWII twin registry was limited to white males led some researchers to conceive of the Black Elderly Twin Study, or BETS. These researchers approached the Social Security Administration with the idea that twins could be identified in SSA administrative records. Specifically, the electronic file of applications for a social security card, known as the NUMIDENT when sequenced by social security number and as the ALPHIDENT when sequenced by name, contain the following personal identifiers useful for assembling twin pairs: surname, date of birth, race, mother"s full name, father"s full name, and city and State or country of birth. Furthermore, often the social security numbers of twins are similar; in particular, if the twins applied at the same time, their social security numbers may have been issued consecutively.

The Social Security Administration was unable to act favorably on this request because of the confidentiality of SSA records. The Health Care Financing Administration (HCFA), now the Centers for Medicare and Medicaid Services, has in its charter a commitment to cooperate in health-related research initiatives, and was able and willing to try to

pair records in its master file of enrollees in Medicare. However, while these records contain surname, date of birth, and race, they do not contain other items captured on the application for a social security card which are important for matching – parents‟ names and place of birth. Because of the significance of the "false positive" problem, the BETS research team was limited to a fairly small set of twin pairs that they were confident of, and the project fell far short of expectations.

In recent years the Social Security Administration has been publishing annual lists of the most common baby (given) names. Because of the Enumeration at Birth process, the names of the very great majority of infants born in a year are known to SSA within a short while. The "baby names" webpage is one of the most popular SSA web pages, and SSA exploits the popularity of the page to make available to the page‟s visitors information about the social security program that parents and guardians should know. The site also provides lists of the most popular name pairs for twins! We do this by matching records for the application for a social security card on surname, date of birth, and State of birth to get candidate pairs. Then we compute a composite score with points given for agreement on mother‟s name, on father‟s name, on city of birth, on race, and on how close the pair of social security numbers are to each other. If the composite score for a candidate pair equals or exceeds a certain threshold, the pair is declared a twin pair. We also find triplets, quadruplets, quintuplets, and sextuplets, but these don‟t go on the website. (For 2009, for example, we found 127,116 twins, 5,307 triplets, 284 quadruplets, 40 quintuplets, and 6 sixtuplets. By comparison, the National Center for Health Statistics reports for 2009 137,217 twins, 5,905 triplets, 355 quadruplets, and 80 infants of plurality 5 or more.)

How often are the ssns assigned to twins close to each other? In 2005 I was able to obtain from the National Academy of Sciences records from its World War II twin registry of persons known to be deceased at that time – about 17 thousand persons. Included were records for over 5 thousand pairs in which both members were known to be deceased, among which were 4,449 pairs for which the record contained social security numbers for both members. The ssns were consecutive 14 percent of the time, and were different by between 2 and 11 another 2 percent of the time. (There is an interesting zygosity differential: the numbers are consecutive for 20 percent of the identical twins, for 10 percent of the fraternal twins, and for 13 percent of those with zygosity unknown.)

By comparison, because of Enumeration at Birth the percentage of infant twins in recent years with "consecutive" numbers is very high; among 2010 twin births, for example, two-thirds of the time the twins were issued consecutive numbers. As mentioned, the proximity of the social security numbers issued to twins helps us to identify twin pairs: we estimate that because of the randomization in ssn issuance which began 7 months ago, about one-sixth of the pairs whose composite score reached the threshold will no longer do so and thus will not be identified as twins with the current algorithm.

Indochinese refugees

The old structure of the social security number allowed for the identification of certain special groups; once a group is defined in terms of ranges of social security numbers, the richness of the information in the administrative records of the Social Security Administration can be mined to paint a statistical portrait of the group. Most obviously, persons with numbers beginning with 700 to 728 are railroad employees issued numbers by the Railroad Retirement Board through June 1963. The special group that I discuss here, however, is the set of Indochinese refugees to whom were issued ssns in certain pre-determined number ranges. Not all Indochinese refugees were issued these numbers, but those who were comprise a significant subset of that population.

Saigon fell in April 1975. About 130 thousand Indochinese refugees arrived in the United States between June and December of that year (Marsh 1980), and about 123 thousand refugees were issued in that same year a social security number in the pre-determined number ranges. Almost half a million more Indochinese refugees arrived here during the following 8 years, but only about 88 thousand of these have an ssn in the special ranges. Because those issued an ssn in the special ranges in 1975 are almost all of the 1975 Indochinese refugees coming to the United States in that year, we limit our analysis to this entry cohort, and the results we present pertain to members of that cohort with the special ssns. This will also provide us with a more homogeneous group; for example, these earlier Indochinese refugees were wealthier and better-educated than those who came later.

The 1975 entry cohort is young overall, and men somewhat outnumber women, as the table below shows. Social Security Administration records provide a broad array of information on the group, including information on citizenship, assimilation, earnings history, and geography.

Age in 1975	Males	Females	Unknown	Total
Less than 5	8000	7532	47	15579
5-14	16503	14841	86	31430
15-24	18001	13534	102	31637
25-34	11865	9256	51	21172
35-44	6162	5317	21	11500
45-54	3501	3061	7	6569
55-64	1438	1896	15	3349
65-74	581	940	3	1524
75-84	155	261	2	418
85 and over	28	72	0	100
Total	66234	56710	334	123278

The application for a social security card solicits information on country of birth. This is South Vietnam for the great majority of refugees in this entry cohort (107,814), with smaller numbers born in North Vietnam (7,803), Cambodia (4,719), and Laos (1,179).

Beginning May 1981 the application for a social security card solicits information on citizenship. Although our population originally applied for a social security number in 1975, many have since filed another application either because some of the original information changed or because of the need for a replacement social security card. (Re-application is more prevalent among women than among men, presumably because of the need to report a surname change upon marriage.) Among the approximately 74 thousand persons in our population who have filed more than one application and the relevant status is coded on the re-application, 86 percent self-identified as citizens.

The surnames are very distinctly "Vietnamese" among the 116 thousand persons in our cohort who were born in Vietnam, South or North. Almost one-third are surnamed NGUYEN. Other common surnames are TRAN (11%), PHAM (7%), and LE (6%). Overall, 81% of the surnames are on a list developed by Lauderdale and Kestenbaum (2000) of 95 distinctive and common Vietnamese surnames.

The extent of changes over time in surnames should be correlated with the extent to which the refugees may be marrying outside their racial group and otherwise assimilating into their new environment. Among the approximately 20 thousand Vietnamese refugee girls under age 15 in 1975, 82 percent had a distinctively Vietnamese surname on their original social security number card but only 62 percent had a distinctively Vietnamese surname on their latest social security number card. For the subset of girls under age 5 in 1975, the decrease was from 80 percent to 56 percent.

How do the mature earnings of the men in the 1975 refugee entry cohort compare with the mature earnings of all U.S. men born in the same period? For the cohort ages 25-34 upon entry in 1975 we look at earnings in 1995, and

for the cohort ages 15-24 upon entry in 1975 we look at earnings in 2005, when the average age of the cohorts is about 50. The older male cohort, those ages 25-34 upon entry in 1975, had median earnings 20 years later at average age 50 about the same as the median earnings for all U.S. male earners in that birth cohort ($33,063 vs. $32,918), and a smaller fraction had earnings at or above the maximum taxable for social security purposes ($61,200 in 1995), 12% vs. 18%. However, the younger male cohort, those ages 15-24 upon entry in 1975, earned more 30 years later at average age 50 than their counterparts: the median earnings are $47,429 and $40,961, respectively, for our group and for all U.S. male earners of comparable age, and the respective percentages earning at least the maximum taxable in 2005 (of $90,000) are 18 percent and 15 percent. These comparisons use earnings subject to social security taxes and are limited to men with some earnings in the reference year.

Although the Indochinese refugee entry cohort of 1975 was a young population at entry, by mid-2009, 34 years later, there were records in the Social Security Administration‟s Master Beneficiary Record file for a large number of cohort members. About 28 thousand were, or had been, entitled to social security monthly benefits; another 8 thousand were, or had been, enrolled in the Medicare program; and another 3 thousand were denied social security benefits and were not enrolled in Medicare. The Master Beneficiary Record contains the current or last address of these 38 thousand cohort members, and we can easily determine that more than half reside in either California or Texas. About 16 thousand reside in California, including 6 thousand in Orange County, and about 6 thousand reside in Texas, including 3 thousand in Harris County (Houston area).

Conclusion

We have illustrated some of the research capability which will be compromised by the new ssn-assignment methodology. We do not doubt that randomization is necessary, but are merely pointing out the diminished functionality of the social security number.

References

Acquisti, Alessandro and Ralph Gross. 2009. *Predicting Social Security Numbers from Public Data.* Proceedings of the National Academy of Sciences 106(27): 10975-10980.

Block, Gladys, Genevieve M. Matanoski, and Raymond S. Seltser.1983. *A Method for Estimating Year of Birth Using Social Security Number*. American Journal of Epidemiology118: 377–395.

Jablon, Seymour, James Neel, Henry Gershowitz, and Glenn Atkinson. 1967. *The NAS-NRC Panel: Methods of Construction of the Panel, Zygosity Diagnosis, and Proposed Use.* American Journal of Human Genetics 19(2): 133-161.

Lauderdale, Diane S. and Kestenbaum, Bert. 2000. *Asian American Race Identification by Surname.* Population Research and Policy Review 19: 283-300.

Marsh, Robert E. 1980. *Socioeconomic Status of Indochinese Refugees in the United States: Progress and Problems.* Social Security Bulletin 43(10): 11-20.