Should Social Security numbers be replaced by modern, more secure identifiers?

William E. Winkler[1]
Statistical Research Division, Bureau of the Census, Washington, DC 20233

Social Security numbers (SSNs) were originally developed as unique identifiers well before recent times when sophisticated methods of using and combining data files became available. Although the main legislated uses of SSNs are by the Social Security Administration (SSA) and the Internal Revenue Service (IRS), SSNs have become ubiquitous as identifiers in both credit files and even health-related files. With many files, the SSN, along with other identifying information such as name, address, telephone number, and date-of-birth, is the primary means of corroborating that an individual from one data source is the same individual in another data source (1).

In our credit-driven society, individuals often want their applications for new credit to be approved as quickly as possible. Approval is often accomplished by comparing information on the credit application with an appropriately designed external database. Speed of linkage may be improved by using only SSN and date-of-birth. Because typographical error is common (say, from keying a hand-written form), the linkage procedures may only require that 7 of 9 characters in the SSN agree and the components of date-of-birth (day-of-birth, month-of-birth, and year-of-birth) approximately agree. The linkage may also use names in a procedure that accounts for minor typographical error. If name is not used in the linkage, then, as noted by Acquisti and Gross (1) in this issue of PNAS, an identify thief can use a new name and mailing address along with the "verified" SSN–date-of-birth combination to obtain new credit.

Acquisti and Gross (1) demonstrate that it is possible to predict SSNs for a moderately large proportion of the population. This is particularly true for individuals who received SSNs via the Enumeration-At-Birth (EAB) procedure that began in 1993. The prediction models are greatly facilitated by SSA's own documented procedures for creating SSNs and publicly available SSA Death Master File (DMF) information that was intended to help prevent fraud and identify theft. To clarify and provide a precursor to later arguments, I repeat some of the description of Acquisti and Gross. The first 5 digits of the SSN are assigned geographically with certain states getting known sets of digits. The precise ordering and specific values have been available in public documents for years (2, 3). The first five digits are assigned in known order (not consecutive) and, within each set of the specific values of the first five digits, the last four digits are assigned consecutively from 0001 to 9999. Although SSA documentation (2, 3) specifically states that the last four digits of SSN are randomized, Acquisti and Gross disprove that valid randomization occurs.

The Acquisti–Gross procedures (1) allow them to predict the first five digits of the SSN with high accuracy. Acquisti and Gross refine their model using the DMF information about the patterns present in the SSNs and dates-of-birth. With the refined model, it is possible to predict the last four digits with accuracy within a range of 100 for individuals born in 1993 or later. The accuracy is much lower for other years. Many of the web-based "identity-verifying" sources allow typographical error in the SSN and a number of verification queries up to a fixed upper bound from a given computer. By varying the guess of the SSN in a range of 100 (or even greater ranges) and using queries from several computers, it is possible to verify a given combination of SSN and date-of-birth to compromise the identity of an individual.

The SSN is not a secure identifier, particularly for individuals born in 1993 and later.

The main issue is that, as Acquisti and Gross demonstrate (1), the SSN is not a secure identifier, particularly for individuals born in 1993 and later. If the SSN is not secure in the sense that it is straightforward to associate it with an individual for whom a name and date-of-birth are available, then it can be very easy to steal such an individual's identity.

Having one's identity stolen can be exceptionally costly (1–3 years, $30,000 or more in expenses) (4–6).

Modern computer environments and capabilities necessitate a secure, accurate, unique, and verifiable identifier. I suggest three changes to existing SSN-assignment procedures that are reasonably straightforward to implement and that may serve as a precursor to more appropriate procedures. The first is to use a different random ordering of the last four digits of the SSN within each group as determined by the first five digits of the SSN. This straightforward change does not affect any of the subsequent legitimate uses of the SSN and should be implemented. Equally easy to implement and even more secure would be for SSA to issue SSNs in a given state at random from the entire set of remaining SSNs that are available to the state. The second change is to add a check digit as an extra field in the SSN (7). Check digits ensure that a set of integers are keyed correctly 90% of the time. The procedure does this by computing a verifying check digit from the existing 9 digits that must agree with the keyed or available check. The "check digit" can be stored in a separate location although, ideally, it might be stored in a location that is adjacent to the SSN. If two check digits were used, then it would be possible to ensure that 99% of SSNs were keyed correctly.

The third change would be to add a pair of digits to deal with the vintage of SSNs. The current 9-digit SSN does not have sufficient numbers for 300+ million Americans, deceased individuals, and others such as certain foreign nationals who need SSNs as part of their U.S. employment. The pair "00" might be associated with most current SSNs and could again be stored in a nonadjacent location. Because some individuals already have two assigned SSNs (8), identical SSNs are sometimes assigned to different individuals, and some geographic regions may be close to running out of SSNs, the SSA could use "01," "02," and so on to disambiguate other sets of SSNs. The third change seems

Author contributions: W.E.W. wrote the paper.

The author declares no conflict of interest.

See companion article on page 10975.

[1] To whom correspondence should be addressed at: Statistical Research Division, Bureau of the Census, 4600 Silver Hill Road, Washington, DC 20233-9100. E-mail: william.e.winkler@census.gov.

crucial because SSA will possibly be running out of sufficient, unassigned SSNs within 70 years (9). The second change facilitates verifying that a transcribed/keyed SSN agrees with SSA's main Numident database containing all verified SSNs and associated information. A third-party group (or individual) with suitable expertise would need to verify that the SSA procedures were properly implemented.

There are two questions related to the general privacy of individuals. First, will SSA be able to issue new, replacement SSNs to individuals from 1993 until the time when SSA implements more secure procedures? Many individuals born from 1993 have significantly increased risk of identity theft. Second, will the credit-granting industry and other groups that need to verify identities adopt procedures that somehow significantly reduce the possibility of identity theft for most individuals? Because millions of individuals are affected by identity theft annually (4–6), the ease with which identity-verifying procedures are compromised needs to be reduced.

1. Acquisti A, Gross R (2009) Predicting Social Security numbers from public data. *Proc Natl Acad Sci USA* 106:10975–10980.
2. Social Security Administration (undated) *SSA's Program Operations Manual System*, https://5044a90.ssa.gov/apps10/poms.nsf/.
3. Jabine TB (1985) Properties of the Social Security number relevant to its use in record linkages, *Record Linkage Techniques 1985*, eds Alvey W, Kilss B (Department of the Treasury, Internal Revenue Service, Washington, DC), pp 219–225. Available at www.fcsm.gov/working-papers/RLT_1985.html.
4. Claburn T (2007) Identify theft: Costs more, tech less. *InformationWeek*, October 22, article ID 202600312.
5. Rubenking J (2004) Identity theft: What, me worry? *PC Magazine*, March 2. Available at www.pcmag.com/article2/0,1759,1522469,00.asp.
6. Burger AK (2008) The cost of identity theft: Part 1, Beyond dollars and cents. *E-Commerce Times*, February 5. Available at www.ecommercetimes.com/story/61515.html.
7. Herzog TA, Scheuren F, Winkler WE (2007) *Data Quality and Record Linkage Techniques* (Springer, New York).
8. Social Security Administration (undated) Why are there multiple Social Security numbers on my statement?, http://ssa-custhelp.ssa.gov/cgi-bin/ssa.cfg/php/enduser/std_adp.php?p_faqid=120&p_created=955568058.
9. Barnhart JAB (2006) Written answers by the Commissioner of the Social Security Administration to questions from J McCrery, Chairman of the Subcommittee on Social Security, March 16. Available at http://waysandmeans.house.gov/hearings.asp?formmode=view&id=4979&keywords=Barnhart+McCrery+March+16.