I Just Found 10 Million SSNs

Alessandro Acquisti,[1*] Ralph Gross,[2]

[1]H. John Heinz III School of Public Policy and Management, Carnegie Mellon University,
4800 Forbes Avenue, Pittsburgh, 15213, USA
[2]School of Computer Science, Carnegie Mellon University,
4800 Forbes Avenue, Pittsburgh, 15213, USA

[*]To be presented at BlackHat 2009.

Introduction. In 1984, Perrow noted that critical failures are likely outcomes for complex systems whose parts interact in unpredictable ways (*1*). Increasingly complex information systems also can lead to unexpected failures: we show that interactions across data sources can be exploited to predict private SSNs from public information (*2*).

The threat originates from the interaction of three (individually innocuous) trends: greater (self)publication of personal information; well-meaning government attempts to prevent SSN fraud (which backfire); and the increasing automation of SSN assignment systems (which introduces regularities attackers can exploit). An attacker could exploit these trends by analyzing publicly available records from the SSA Death Master File (DMF) to detect statistical patterns in the SSN assignment for individuals whose deaths have been reported to the SSA; and then, by interpolating an alive person's state and date of birth with the patterns detected across deceased individuals' SSNs, to predict a *range* of values likely to include his SSN. Birth data, in turn, can be inferred from several offline and online sources, including voter registration lists or online "white pages" and the profiles millions of individuals publish on social networking sites (*3*).

Methodology and Results. Following the Enumeration at Birth initiative (an anti-fraud program which started extending nationwide in 1989), the overwhelming majority of US newborns obtain their SSNs shortly after birth (*4*). While the assignment process remains inherently noisy, we hypothesized that EAB increased the likelihood that, for US-born applicants, the state of SSN application would be their state of birth, and the date of application would correlate with their birthday. The Death Master File - a publicly available file reporting SSNs, names, dates of birth and death, and states of SSN application for individuals whose deaths were reported to the SSA.[1] - contained the data necessary to verify such regularities in the assignment scheme, and then predict individual SSNs based on the dates and states of birth of their applicants. Specifically, we used the DMF as an *analysis set* to identify assignment patterns and as a *test set* to test the accuracy of SSN predictions based on extrapolated patterns. We selected two success metrics: whether we could identify with one attempt an SSN's first five digits (as the last four may be discerned elsewhere) and the entire SSN in fewer than 1,000 attempts (making SSNs akin to 3-digit PINs). We predicted each DMF record's first five digits (its AN and GN) based on the *most frequent* ANs and GNs assigned to DMF records with the same state of application and born around the target record's birthday; we predicted its last four digits (its SN) combining its birthday with coefficients estimated from linear regressions over individuals' SNs with similar birthdays and same state of application as the target. As hypothesized, we found widespread increases in prediction accuracies after 1989 (the onset of the nationwide EAB program), particularly for less populous states (fewer daily births determine more discernable patterns). For instance, we accurately predicted the first five digits of only 2% of California records with 1980 birthdays versus 90% of Vermont records with 1995 birthdays; on average, we matched the first five digits for 44% of all records born nationwide after 1988. Imagining a brute force algorithm where, for each target, the attacker tries out the predicted SSN before

[1] `http://www.ntis.gov/products/ssa-dmf.aspx`.

moving up and down the SNs in 1-integer steps for the following attempts (while keeping the predicted ANs/GNs constant), fewer than 1,000 attempts may be sufficient to identify the SSNs of 8.5% of all individuals born after 1988.

Implications. When fewer than 1,000 attempts are sufficient to predict massive amounts of SSNs, various brute force attacks become economically feasible: an attacker could try to identify an SSN by testing subsets of variations predicted by the algorithm across different channels, including phishing emails, online instant credit approval services, or the SSA's own SSN Verification Service[2] and the Department of Homeland Security's E-Verify system.[3] Unlike traditional identity theft strategies, these channels allow attackers to test, covertly and cheaply, multiple variations of predicted SSNs for massive numbers of targets, while choosing them based on demographic traits, bridging the gap between statistical predictions and actual identity theft.

In order to prevent damages deriving from the predictability of SSNs, various parties may consider mitigating strategies: the SSA can cheaply randomize the assignment scheme (eliminating identification risks for new SSNs); social networking sites can discourage the revelation of birth data by tuning default privacy settings; other services - from which birth information may be inferred - can reassess data accessibility policies. None of the above strategies, however, is foolproof, and all carry unintended consequences. Furthermore, assigned SSNs cannot be revoked to avoid future frauds, and exposed data cannot be taken back. Industry and policy makers may need, instead, to reassess our reliance on SSNs as passwords. They may further scrutinize the security of the SSN Verification Service and E-Verify, and the recommendations of the President's Identity Theft Task Force, which focused on limiting "unnecessary" uses of SSNs while preserving their role as integral part of the financial system (*5*) (under the presump-

[2] `http://www.ssa.gov/employer/ssnv.htm`.

[3] `http://www.dhs.gov/xprevprot/programs/gc_1185221678150.shtm`.

tion that SSNs can, in fact, remain private). Policy debate may further evaluate various states' recent legislative initiatives to protect SSNs by restricting the public usage of *only* their first five digits:[4] considering how predictable we found those digits to be, such legislations may be misguided.

References

1. C. Perrow, *Normal Accidents: Living with High-Risk Technologies*. Basic Books, 1984.

2. A. Acquisti and R. Gross, "Predicting social security numbers from public data," *Proceedings of the National Academy of Science*, Forthcoming, 2009.

3. R. Gross and A. Acquisti, "Privacy and information revelation in online social networks," in *Proceedings of the ACM Workshop on Privacy in the Electronic Society*, pp. 71–80, ACM, 2005.

4. Social Security Administration, "Report to Congress on options for enhancing the social security card," 1997. http://www.ssa.gov/history/reports/ssnreport.html.

5. The President's Identity Theft Task Force, "Combating identity theft: A strategic plan," 2007. http://www.idtheft.gov/reports/StrategicPlan.pdf.

[4] `http://www.ncsl.org/programs/lis/privacy/SSN2007.htm`.