



### DEATH MASTER FILE

Acquisti and Gross's study relied on the Social Security Administration's Death Master File (DMF), a public database of SSNs along with birth and death dates and states of birth for every deceased Social Security beneficiary. The purpose is to prevent impostors from using SSNs of non-living people. Using sophisticated statistical analysis, the researchers found they could detect patterns in the DMF data that made it possible to predict SSNs of living people.

When combined with a living person's date and state of birth, these patterns can significantly narrow

the possibilities of pinpointing that person's SSN. Since place and date of birth information is available from many sources — commercial databases, public records (including voter registration lists), and millions of profiles on social networks, personal web sites and blogs — the possibility to predict SSNs becomes a reality.

The statistical patterns and birth information can predict SSNs because the Social Security Administration's methods for assigning numbers, based partly on geography, are well known. For most individuals born since 1989, furthermore, SSNs are assigned shortly after birth, making those numbers easier to predict than for earlier birth years.

Acquisti and Gross tested their method, a complex statistical algorithm, on DMF records for people who died between 1973 and 2003. For people born after 1988, they could with a single attempt identify the first five digits for 44 percent of people. For people born between 1973 and 1988, they could identify the first five digits in 7 percent of the cases. With more attempts, up to but fewer than 1,000, they were able to identify all nine SSN digits for 8.5 percent of people born after 1988.

For smaller states and recent years of birth, their accuracy was higher than for large states and earlier years. They needed 10 or fewer attempts, for instance, to predict all nine digits for one out of 20 SSNs issued in Delaware in 1996. "If you can successfully identify all nine digits of an SSN in fewer than 10 or even fewer than 1,000 attempts," said Acquisti and Gross, "that Social Security number is no more secure than a three-digit PIN."

When the researchers tested their method on non-DMF data — using birth dates and hometowns that students had self-reported on popular social networking sites — the results were almost as good despite the typical inaccuracies of social-network data. The researchers used enrollment records to confirm the accuracy of their predictions, though they didn't

confirm any individual SSNs, only aggregate measures of accuracy.

"Dramatically reducing the range of values wherein an individual's Social Security number is likely to fall makes identity

theft easier," says Gross. A fraudster who knows just the first five digits might use a phishing e-mail to trick the person into revealing the last four digits.

Or a fraudster could use networks of compromised computers, or "botnets," to repeatedly apply for credit cards in a person's name until hitting the correct nine-digit sequence.

### A BIG SHOVEL FOR MASSIVE DATA

It would have been difficult, if not impossible, to obtain these findings, says Acquisti, without high-performance computing resources such as PSC's Pople. After first working with desktop computers and coming to a bottleneck in their work, the researchers approached PSC. This was mid-2008, at an important phase in the project. "At that stage," said Acquisti, "we had a rough idea of the results, but to go forward we had to try many different variations of the algorithms. It would have been incredibly difficult to do this, or taken much, much longer without access to this system."

Acquisti and Gross and several graduate students who worked with them turned to Pople. Named for Nobel laureate chemist John Pople of Carnegie Mellon), this system features 768 cores (processors) and 1.5 terabytes of shared memory (all of memory accessible from each core). Working with a core dataset of about eight gigabytes, the researchers used 100 processors for up to eight hours for each of a series of seven runs.

PSC staff installed Octave — an open-source version of the programming language MATLAB

**"GIVEN THE INHERENT VULNERABILITY OF SOCIAL SECURITY NUMBERS, IT IS TIME TO STOP USING THEM FOR VERIFYING IDENTITIES AND REDIRECT OUR EFFORTS TOWARD SECURE, PRIVACY-PRESERVING AUTHENTICATION METHODS."**

— and wrote a script to submit a large number of parallel Octave jobs simultaneously. This facilitated the Acquisti team's interactive process, which involved doing many runs representing different states and computational strategies, checking and analyzing results and re-thinking before running more variations. PSC's consulting, said Acquisti, was "extremely helpful."

"This project," said Sergiu Sanielevici, PSC director of scientific applications and user support, who also leads user support and services for the TeraGrid, "exemplifies how powerful systems like Pople can open doors to data-mining and data-centric research in fields not traditionally associated with HPC, such as the social sciences, and make it possible to get answers that would otherwise be impractical or impossible."

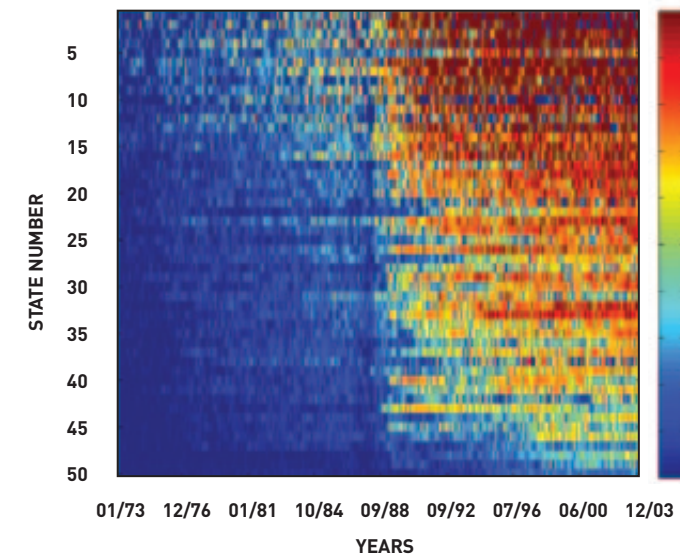
Carnegie Mellon graduate students Jimin Lee, Ihn Aee Choi, Dhruv Deepan Mohindra, and Ioanis Alexander Biternas Wischniensi collaborated in this research and did much of the hands-on computational work.

Future SSNs could be made more secure, say Acquisti and Gross, by switching to a randomized assignment scheme. Protecting people who already have been issued numbers, however, is more difficult. Given the ease with which SSNs can be predicted — particularly the first five digits and particularly for the millions of Americans born since 1988 — legislative and policy initiatives aimed at removing the numbers from public exposure, or redacting the first five digits, added Acquisti, may be well-meaning but misguided.

"Given the inherent vulnerability of Social Security numbers," says Acquisti, "it is time to stop using them for verifying identities and redirect our efforts toward implementing secure, privacy-preserving authentication methods." Methods to consider include two-factor authentication, similar to the PIN number-card combinations used for bank accounts, and digital certificates.

### MORE INFORMATION

[www.psc.edu/science/2009/privacy/](http://www.psc.edu/science/2009/privacy/)



**PREDICTION SUCCESS RATIOS**  
Prediction accuracies for Death Master File records with January 1973 to December 2003 birthdays across the 50 states. Ratios of accurate prediction for the first five digits (top), and ratios of accurate prediction for the complete SSN with less than 1,000 attempts (bottom). In each quadrant, columns represent months, and rows represent states (sorted by their 1973 births, lowest to highest). The colors in each cell represent ratios (from 0 through 1, dark blue through red) out of monthly SSN counts. This fairly unassuming graphical figure, notes Acquisti, represents results of "more than 700,000 regressions over a very large data set."

