Predicting Social Security Numbers from Public Data

Alessandro Acquisti and Ralph Gross

Heinz College/CyLab
Carnegie Mellon University

Research support from National Science Foundation, U.S. Army Research Office (through CyLab), Carnegie Mellon Berkman Fund, and Pittsburgh Supercomputing Center

Black Hat USA 2009

Overview

1. Show that Social Security numbers (SSNs) are predictable from publicly available data
   - Knowledge of an individual's birthday and birthplace can be exploited to infer narrow ranges of values likely to include that individual's SSN
   - This is due in part to well-meaning, but counter-effective, public policy initiatives
2. Highlight associated risks and implications
3. Discuss possible risk-mitigating strategies & policies

Social Security numbers: Identifiers vs. Authenticators

- SSNs were designed and issued by the Social Security Administration (SSA) for the first time in 1936 as identifiers for accounts tracking individual earnings
- Unfortunately, over time they started being used, and abused, as authentication devices
  - Notwithstanding warnings by SSA, FCT, GAO, scholars, and so forth
  - Naturally, the same number can't be used securely both as identifier and for authentication

From SSNs to identity theft

- The wide availability of SSNs, and their dual use as identifiers and authenticators, make identity theft easy and widespread
- Knowledge of somebody's name, DOB, and SSN is often **sufficient condition** for access to financial, medical, and other services
  - Sometimes, even applications with just 7 out of 9 correct digits are accepted as valid (FTC 2004)

The assignment scheme of SSNs is public knowledge (we did not break any code!)

- Each SSN has 9 digits:
  - XXX-YY-ZZZZ
- ... and is composed of three parts:
  - Area number: XXX
  - Group number: YY
  - Serial number: ZZZZ
- The SSN issuance scheme is complex, but not stochastic
  - The SSA itself has for a long time publicly revealed its details

The scheme follows geographical and chronological patterns

- This is well known
  - In fact, inference of the likely time and location of SSN applications based on their digits has been exploited to catch fraudsters and impostors
- However, the SSA also states that the SSN assignment process is, effectively, random:
  - *"SSNs are assigned **randomly** by computer within the confines of the area numbers allocated to a particular state based on data keyed to the Modernized Enumeration System" (RM00201.060)*

Chances of correctly matching SSN digits by random guess, under *status quo* knowledge

	Alaska		New York	
	First 5 digits with 1 guess	All 9 digits with < 1,000 guesses	First 5 digits with 1 guess	All 9 digits with < 1,000 guesses
No auxiliary knowledge	0.0014%	0.00014%	0.0014%	0.00014%
Knowledge of state of SSN application	1%	0.1%	0.012%	0.0012%

However: Reasons to believe that the assignment lacks sufficient randomness

- In the last 30 years, SSN issuance has become more regular
  - Increasing computerization of the public administration, including SSA and its various fields offices
  - After 1972, SSN assignment centralized from Baltimore
  - Tax Reform Act of 1986 (P.L. 99-514)
  - After 1989, Enumeration at Birth Process (EAB)
    - Prior to 1989, only small percentage of people received SSN when they were born
    - Currently at least 90 percent of all newborns receive SSN via EAB together with birth certificate

Hence, two hypotheses

1. We expected SSN issuance patterns to have become more regular over the years, i.e. increasingly correlated with an individual's birthday and birthplace
   - This should be detected through analysis of available SSN data
2. We expected these patterns to have become so regular that it is possible to infer *unknown SSNs* based on the patterns detected on *available SSNs*
   - This should be verified by contrasting estimated SSNs against known SSNs

Compared to previous work

- Outside the SSA, the current understanding of the assignment of the first 3 digits was incorrect, and the relationship between demographic patterns and the sequentiality of the last 4 digits was unexplored
  - Hence, previous work in this area focused on **inferring the likely year or years and state of SSN issuance of a <u>known</u> SSN** (e.g., [Wessmiller, 2002], [Sweeney, 2004], [EPIC, 2008])
- We focused on the inverse, harder, and much more consequential inference: **exploiting the presumptive day and location of SSN application to predict <u>unknown</u> SSNs**

Chances of correctly matching SSN digits by random guess, under our algorithm

	Alaska, 1998		New York, 1998	
	First 5 digits with 1 guess	All 9 digits with < 1,000 guesses	First 5 digits with 1 guess	All 9 digits with < 1,000 guesses
No auxiliary knowledge	0.0014%	0.00014%	0.0014%	0.00014%
Knowledge of state of SSN application	1%	0.1%	0.012%	0.0012%
Predictions based on our algorithm	**94%**	**58%**	**30%**	**3%**

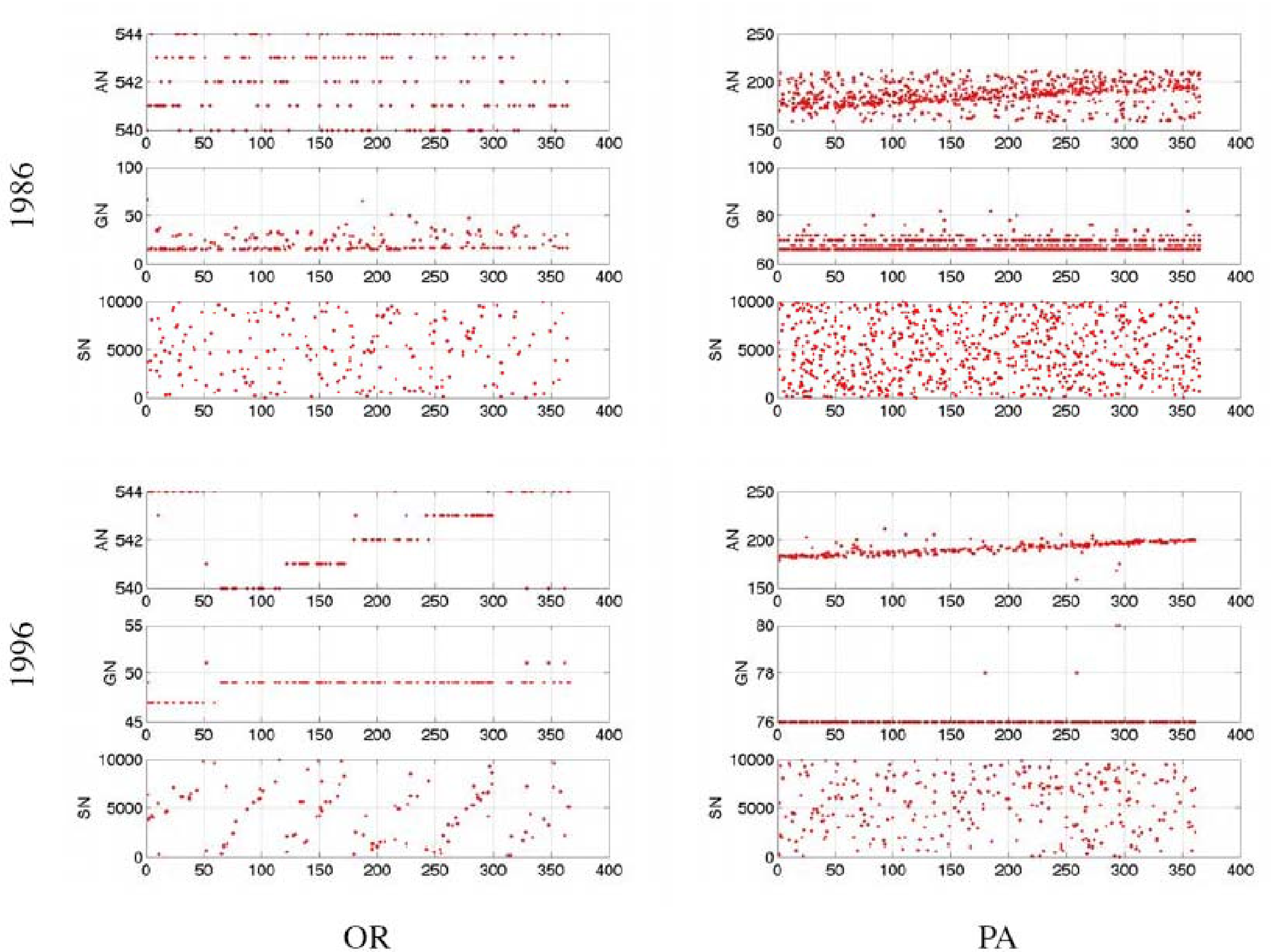
The Death Master File

- The Social Security Administration's Death Master File is a publicly available database of the SSNs of individuals who are deceased
  - One of the purposes of making this data available was to combat fraud
  - Unfortunately, it can also be analyzed to find patterns in the SSN issuance scheme
- We used DMF data to find patterns in the issuance of SSNs by date of birth and State of SSN issuance for deceased individuals
  - Namely, we sorted records by reported DOB and grouped them by reported State of issuance
  - An iterative process

A DMF record (example)

Name	Birth	Death	Last Residence	SSN	Issued
JOHN SMITH	21 Jun 1904	Oct 1979	33540 (Zephyrhills, Pasco, FL)	022-10-3459	Massachusetts

SSN assignment patterns: Two representative States



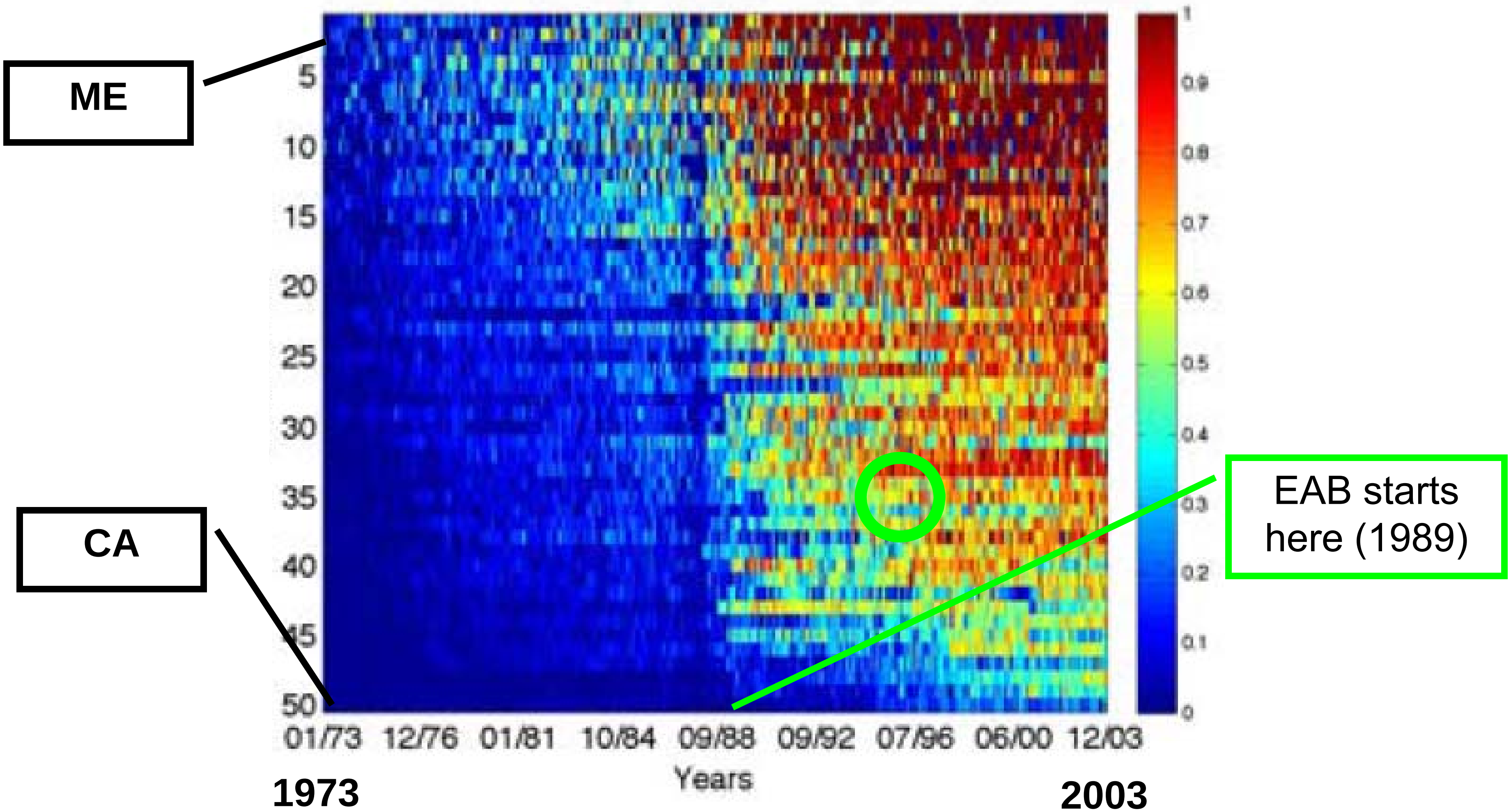
SSN predictions

1. TEST 1: We used more than half a million DMF records to detect **patterns in SSN issuance based on birthplace and state of issuance**, and used those patterns to predict (and verify) individual SSNs in the DMF
2. TEST 2: We mined data from an online social network to retrieve **individuals' self reported birthdays and birthplaces, and estimated their SSNs by interpolating that data with DMF patterns**. We verified the estimates using official Enrollment data using a protected (and IRB approved) protocol

Two “success“ metrics

1. Whether we could predict the first 5 digits of an individual’s SSN with one attempt
2. Whether we could predict the entire SSN with fewer than 10, 100, and 1,000 attempts
   - Note: 1,000 attempts is equivalent to 3-digit PIN
   - That is, very insecure and vulnerable to brute force attacks

Test 1: AN-GN predictabilty (first 5 digits)



Test 1: Overall results for DMF records

- With a single attempt (<u>first five digits only</u>):
  - 7% (1973- 1988)
  - 44% (1989-2003)
- With 10 attempts (<u>complete 9-digit SSNs</u>):
  - 0.01% of (1973- 1988)
  - 0.1% (1989-2003)
- With 1,000 attempts (<u>complete 9-digit SSNs</u>):
  - 0.8% (1973-1988)
  - 8.5% (1989- 2003)

- These are weighted averages – for smaller states and recent years, prediction rates are higher. E.g., **1 out of 20 SSNs in DE, 1996, are identifiable with 10 or fewer attempts**

Test 2: From social networks data to SSNs

- In Test 2 we used birthday data of 621 *alive* individuals to predict their SSN, based on interpolation with DMF data
  - Our sample: born in 1986-1990 (i.e., mostly before EAB)
  - In most populous states (i.e., worst case scenario)
- Birthday and birthplace data can be obtained from several sources, but most easily, and in mass amounts, from online social networks
  - It is trivial for an attacker to write scripts to penetrate OSN communities and download massive amounts of data

Results and extrapolations

- Test 2 confirmed results of Test 1 (for same mix of years/states of birth)
- **This validates that interpolation of SSN data for deceased individuals and birthday data for alive individuals can lead to the prediction of the latter's SSNs**
- Extrapolating to the US living population, this would imply the identification of around 40 million SSNs' first 5 digits and almost 8 million individuals' complete SSNs
  - *Caveat: Assuming knowledge of birth data!*

Where does birth data come from?

- Personal knowledge
- Online social networks
- Voter registration lists
- Free online people search services
- Commercial databases

From statistical predictions to identity theft

- Statistical predictions do not amount, alone, do identity theft
  - How can you "test" 10, 100, or 1,000 variations of an SSN without raising red flags?
  - *Using botnets and distributed online services for brute force verification attacks*
    - "Tumbling" attacks have been documented by ID Analytics

Verification attacks

- Phishing
- SSNVS: SSN Verification Service (SSA)
- eVerify (DHS)
- Instant credit approval services
  - DOB/SSN match often is **sufficient condition** to get approved for several services

A vulnerable information ecosystem

Availability of birth data
- Commercial databases
- Free online "people" searches
- Voter registration lists
- Online social networks

→ **SSN predictability** → **Distributed attacks**
- Botnets

→ **Online verification systems**
- Instant credit approvals
- eVerify
- SSNVS

→ **SSNs as authenticators**
- CRAs
- Financial institutions
- Medical services
- […]

•Change default settings?
•Change access/security policies?

•<u>Randomize assignment scheme (all digits)?</u>

•Improve personal computer security?

•<u>Be on the alert for distributed attacks?</u>
•<u>Improve real-time coordination? (ID Analytics 2003)</u>
•<u>Improve lax verification procedures?</u>

•<u>Stop using SSNs for authentication, revert to single use as identifiers?</u>

Implications

- Short term
  - Randomize scheme
  - But, this alone not enough:
    - Does not protected issued SSNs; does not resolve authenticator/identifier issue
- Long term
  - Reconsider legislative initiatives focusing on redacting/removing SSNs from documents/public exposure
  - Phase out "authentication" usage
    - "Negligence" argument for businesses that use them as such?
  - "Sunset" solution?
    - E.g., make all SSNs public by year 2014 – transition to secure, private, efficient authentication methods in the meanwhile?

Acknowledgements

Research support from the National Science Foundation under Grant 0713361, from the U.S. Army Research Office under Contract DAAD190210389, from Carnegie Mellon Berkman Development Fund, and from the Pittsburgh Supercomputing Center is gratefully acknowledged

National Science Foundation
WHERE DISCOVERIES BEGIN