Social Insecurity: The Unintended Consequences of Identity Fraud Prevention Policies

Alessandro Acquisti and Ralph Gross

Abstract: Designed as identifiers of accounts tracking US residents' earnings, Social Security numbers (SSNs) have become over time sensitive authenticators for private sector services. Since their abuse is a major vector of identity theft, numerous initiatives have attempted to reduce their public availability. However, recent research has shown that unavailable SSNs may still be accurately predicted from publicly available data. Such predictability may undermine the effectiveness of current strategies aimed at curtailing identity theft. This manuscript examines the policy initiatives that made SSNs predictable, the extent to which their predictability heightens the risks of identity theft for US residents, and the effectiveness of current identity theft prevention strategies in light of said predictability. We find that, surprisingly, a number of past policy initiatives designed to combat identity fraud actually created the conditions for the predictability of SSNs. We also find that current policies aimed at enhancing privacy and security of SSNs, while well-meaning, may too prove countereffective. Our results support alternative identity management solutions, such as banning the usage of SSNs for authentication and replacing them with credential systems based on usable cryptographic protocols.

Keywords: Economics of Information Security; Identity Theft; Social Security Numbers; Online Social Networks.

1 INTRODUCTION

In 1984, Charles Perrow noted that critical failures are likely outcomes for complex systems whose parts interact in unpredictable ways (Perrow 1984). Perrow suggested that attempts to secure such systems can, in fact, precipitate their demise through a chain of unexpected consequences. Increasingly complex information systems can also generate unexpected failures: we show that a combination of anti-fraud policy initiatives, enacted over the last 30 years in the United States, set the conditions for the predictability of Social Security numbers (SSNs) from public data; we also show that such predictability heightens not just statistically, but practically, the risk of identity theft for US residents; and we finally show that current legislative interventions in this area, while well-meaning, may too produce counterintuitive and undesirable consequences.

In the United States, identity fraud often relies on knowledge of the victim's Social Security number. Designed in the 1930s as identifiers of accounts tracking individual earnings (SSA 1996), SSNs became over time the *de facto* authenticator for the American consumer: many services in the private sector take its mere knowledge as a proof of identity (Smith 2002), making SSNs a valuable target for criminals and a prominent vector for identity theft (GAO 2006a).

Numerous policy initiatives have been enacted to prevent identity theft. They include state-level data breach notification laws (Romanosky et al. 2011) and federal legislations such as the Fair and Accurate Credit Report Act of 2003 (FACTA). Many of these initiatives focus on "securing" SSNs: government agencies (including the Social Security Administration (SSA), which issues them) have urged individuals to keep their SSNs safe and confidential (FTC 2006b, SSA 2007b, FTC 2007), and have requested private and public sector entities to limit their public exposure. The premise underlying such strategies is that SSNs can, and should, be *protected*: in other words, by reducing their public availability, we may reduce identity theft, without disrupting

SSNs' role as means of authentication in information services across many industries, and most notably the financial sector (FTC 2006b, SSA 2007b, FTC 2007).

The findings we present in this manuscript cast doubts on the feasibility, as well as opportunity, of such a strategy. Our analysis builds upon the discovery that SSNs may be accurately, albeit only statistically, predicted based entirely on publicly available data (Acquisti and Gross 2009). Such predictability, coupled with the increasing online availability of non-sensitive personal data for massive amounts of US residents, raises the concern that initiatives aimed at hiding SSNs may still leave us at risk of identity theft, because redacted or removed SSNs will remain inferrable from otherwise publicly available information - and, therefore, still vulnerable. Our manuscript investigates these concerns by exploring the policy roots and consequences of the predictability of Social Security numbers. It focuses on three questions:

1. Did identity fraud prevention policies make SSNs predictable from public data (Section 3)?
2. Does the predictability of SSNs beget a practical, or only a theoretical, risk of identity theft (Section 4)?
3. Can current SSN protection policies effectively combat identity theft (Section 5)?

To answer these questions, we employ a combination of empirical tools: fixed-effect regressions, to analyze the prediction accuracies of Social Security numbers vis à vis US demographic trends; an online experiment, to measure the practical predictability of Social Security numbers from real-life data; and numerical simulations, to estimate the effectiveness of various policy initiatives aimed at combating identity theft.

2 BACKGROUND: SSNs AND IDENTITY THEFT

At the heart of the US identity theft problem lies the dual usage of Social Security numbers as both *identifiers* and as *authenticators* (Smith 2002, LoPucki 2003, Solove 2003). Identifiers, as symbols representing an identity, can be made public. For instance, somebody's phone number is an identifier for that individual: the person can share that number with friends, or even publicly. Authenticators, being proofs of identity, are sensitive and private. For instance, the secret personal code a person uses to access her voice mail is a form of authentication. Obviously, the same number should not be used for both purposes - to identify as well as to authenticate an individual. However, this is how Social Security numbers are used in the United States (FTC 2004). Such dual usage of SSNs is one of the major reasons for the high incidence of identity theft in the United States (Solove 2003).

In 1936, following the enactment of the Social Security Act of 1935 (P.L. 74-271), the board that would later become the SSA started issuing SSNs to track earnings and calculate benefits for the newly introduced Social Security program (SSA 1996, RM00201.001). Social Security numbers were never designed for authentication - a practice the Social Security Administration did not initiate, nor encouraged (Smith 2002).[1] Their complex (but not necessarily random) assignment scheme made sense in a world of physical cabinets and printed records (SSA 2011d). As most adult Americans started receiving an SSN, however, Social Security numbers started offering a way to permanently and uniquely identify individuals within and across diverse databases (Ware 1973). Hence, their usage sprawled in a decentralized, unregulated up fashion: different parties (first in the government, then in the private sector) started using SSNs for their own different purposes. SSNs started being collected (and sometimes publicly displayed) by federal, state, and

[1] The SSN card itself, from 1946 till 1972, included a legend on the bottom of the card reading "FOR SOCIAL SECURITY PURPOSES – NOT FOR IDENTIFICATION" (SSA 2011c). We note that the terms "identification" and "authentication" are

local governments or courts; then, they started being requested and stored by banks, telecoms, universities, or healthcare organizations; finally, they started being aggregated, vetted, and traded by information resellers, data brokers, and credit reporting agencies (CRAs).

In 1974, the Privacy Act made it unlawful for governmental agencies to deny rights, benefits, or privileges to individuals who refused to disclose their SSNs. However, the Act did little to curtail their usage in the private sector (Solove 2003). By then, SSNs were on their way to become the primary means of consumer authentication across a host of industries, especially credit and finance (Long 1993, Smith 2000). Today, mere knowledge of someone's SSN and basic personal information has become a necessary and sometimes *sufficient* condition for access to a plethora of services and benefits such as credit cards, mobile phone accounts, or medical services. In most "new account" frauds, criminals only need a victim's name, date of birth, and SSN to create a *new* account in the victim's name (for instance, a new credit card), even without knowledge of the victim's correct address, phone number, or other personal data (Cook 2005, Hoofnagle 2007a, Samuelson Law, Technology & Public Policy Clinic 2007).[2]

2.1 Attempts at Combating Identity Theft by Protecting SSNs

In recent years, the US legislator has attempted to combat identity theft by curtailing the availability of Social Security numbers. Not only have government agencies campaigned to encourage individuals to treat their SSNs as confidential (FTC 2006b, SSA 2007b, FTC 2007), but a number of public policy initiatives have specifically aimed at removing SSNs from public documents and curtailing their trade (GAO 2004, FTC 2006b, SSA 2007b, FTC 2007, GAO 2008).

sometimes used differently across disciplines and scholars. For instance, (Solove 2003) refers by "identification" to the process we instead call here "authentication."

[2] In recent years, massive numbers of SSNs have been accessed by malicious parties due to breaches of corporate and government databases, online scams, or other offline frauds. The growing number of compromised SSNs has been accompanied by alarming rates of identity theft, for an estimated cost to US firms and consumers of $54 billion in 2009 (Greenberg 2010) and $37 billion in 2010 (Ody 2011). While the aggregate losses have decreased between 2009 and 2010, the damage per victim have increased, as the costs of "new-account" frauds (the type we focus on in this manuscript) have risen (Ody 2011).

States such as California, Florida, and Oregon have enacted laws mandating the redaction of SSNs in public records (Office of the Inspector General, SSA 2007, GAO 2008). An even larger number of proposed federal legislative initiatives have aimed at limiting the display or sale of SSNs (such as H.R. 3046 of 2007, H.R. 948 of 2007, S. 238 of 2007, S. 2915 of 2008, and S. 1618 of 2009).[3] Under a November 2008 presidential executive order, federal agencies are no longer mandated to rely exclusively on SSNs as personal identifiers (Executive Order 2008).

While the reduced availability of SSNs and the increased protection of sensitive data are positive trends, neither resolves the underlying vulnerability of a national identity infrastructure centered around employing the same sequence of numbers as identifiers and authenticators. Furthermore, neither trend remedies the threat that even protected, unavailable SSNs, may still be predictable from otherwise publicly available information.

In a recent paper, Acquisti and Gross (2009) showed that all nine digits of an individual's SSN may be statistically, yet often quite accurately, predicted simply based on knowledge of that individual's date and state of birth.[4] Acquisti and Gross (2009) showed that a common interpretation of the assignment scheme held outside SSA (under which only a subset of an SSN digits would be partially predictable) was incorrect, and that even the SSA's description of the assignment process as random (SSA 1996, RM00201.060) overplayed the stochastic components in an otherwise deterministic process. As a result, an inverse inference was possible: predicting *unknown* SSNs from their likely state and time of application.

Using the Death Master File (DMF) – a publicly available database of SSNs, names, dates of birth and death, and states of SSN application of individuals whose deaths have been reported to

[3] At the time of writing (April 2011), none of those bills have become law, possibly due to US legislators' focus, since 2008, on other issues such as healthcare and financial reforms. However, members of Congress can reintroduce bills that did not come up for debate under new numbers in the next session.

[4] While the existence and details of the SSN assignment scheme had been public knowledge for years (Block et al. 1983, Clifton and Marks 1996, Hoofnagle 2007b), that scheme was believed to contain sufficient complexity and noise to render most of an SSN's

the SSA – the authors reported matching, at the first attempt, the first five digits for 7% of all records for individuals born nationwide between 1973 and 1988, and 44% for those born after 1988. With fewer than 1,000 brute-force attempts per target, they identified 8.5% of all 9-digit SSN records with dates of birth between 1989 and 2003 (in a brute force attack, the criminal tries out a number of possible variations of a target's SSN until he hits the correct one, or moves on to the next target; a similar practice is already exploited by cybercriminals and is known as "tumbling" (ID Analytics 2006)). Acquisti and Gross' predictions were more accurate for individuals born in more recent years and in smaller states. For instance, the authors reported identifying 1 out of 20 SSNs in the DMF issued in Delaware in 1996 with just 10 or fewer attempts. As the authors noted, a successful identification of an entire SSN using anything less than 1,000 attempts makes that SSN no more secure than a 3-digit (that is, highly vulnerable) financial PIN (for comparison, ATMs use 4 digits but also require the physical presence of bankcards).

3 DID IDENTITY FRAUD PREVENTION POLICIES MAKE SSNs PREDICTABLE?

Having described the role of SSNs in identity theft, and current public policies aimed at protecting SSNs, the remainder of this manuscript addresses three questions: Did identity fraud prevention policies make SSNs predictable from public data (this section)? Does such predictability beget a practical, or only a theoretical, risk of identity theft (Section 4)? And, can current SSN protection policies effectively combat identity theft (Section 5)?

digits effectively random (see SSA (1996)). Accordingly, previous research in this area merely aimed at predicting the state and time of application of *known* SSNs based on their digits (Sweeney 2004).

3.1 Hypotheses

According to the SSA's Program Operations Manual System (POMS), "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" (SSA 1996, RM00201.060). How can SSNs be randomly assigned, and yet predictable from simple demographic data?

The resolution to the apparent paradox lies in the observation that, while the *assignment* scheme of SSNs is complex, but not random, the actual assignment of individual SSNs depends on the order in which *applications* for SSNs are received by the SSA. Such order used to be, effectively, random – as individuals applied for SSNs at idiosyncratic and unpredictable moments in their lives. However, we conjectured that a number of policy initiatives enacted throughout the 1980s instilled regularities in the time of SSN *application* for an increasing portion of US residents. Specifically, we hypothesized that a set of policy initiatives, paradoxically intended to combat various forms of identity fraud, dramatically increased the likelihood that SSNs would be applied for at a determinate (and therefore inferrable) moment of the applicant' life, making SSNs highly predictable.

First, we considered the Interest and Dividend Tax Compliance Act of 1983 (P.L. 98-67), the Tax Reform Act of 1986 (P.L. 99-514), and the Family Support Act of 1988 (P.L. 100-485). Each of these federal acts enacted included provisions aimed at preventing identity-based tax frauds, by curtailing the number of dependents falsely claimed in tax returns. They did so by requiring parents listing younger dependents in their tax returns to include the dependents' SSNs. For instance, under the Tax Reform Act of 1986, individuals over 5 years of age without an assigned SSN could not be claimed as dependents on tax returns. Two years later, Section 704(a) of the Family Support Act required individuals filing a tax return to include the taxpayer identification number (that is, usually the SSN) of each dependent age 2 or older. Cumulatively, these provisions increased the incentives for parents to apply for their children's SSNs at a young age or, in fact, at birth (Long 1993).

Then, we considered the Social Security Administration's Enumeration at Birth (EAB) program. The percentage of parents who applied for their children's SSNs at birth was increased by the enactment of the EAB in the late 1980s. Under the EAB, parents were allowed to request SSNs for their newborns when the infant's birth was registered by the state (SSA 1997). Among other objectives, the EAB was meant as "a valuable tool in preventing fraudulent acquisition of an SSN and a Social Security card" (Donnelly 1999). Specifically, the program was intended to prevent criminals from obtaining a child's SSN before the child had actually applied for it. The EAB program started in August 1987 in New Mexico, Indiana, and Iowa (Long 1993), and started expanding nationwide a couple of years later. By the end of 1991, 45 states (as well as NYC, DC, and Puerto Rico) had signed agreements to be part of the program (Connecticut, Rhode Island, Oklahoma, Alaska and California were not yet participating at the time). By September 1997, all 50 states had joined the program (SSA 2001). The number of newborns who received SSNs under the EAB kept increasing during the 1990s and after 2000. The SSA estimates that EAB covered approximately 75 percent of all newborns in 1997 (SSA 1997), 90 percent of the Social Security cards issued in 1998 (Donnelly 1999), and 92 percent of SSNs assigned to US citizens by 2005 (Thompson, 1988, SSA, 1996, Office of the Inspector General, SSA 2006).

The 1980s tax initiatives and the Enumeration at Birth program dramatically increased the number of individuals receiving their SSNs at birth. Such effect was cumulative: with each subsequent Act (in 1983, 1986, and 1988) the fraction of US newborns receiving SSNs increased; similarly, the fraction of newborns receiving their SSNs under EAB kept rising throughout the 1990s and 2000s. We hypothesized that these policy initiatives 1) instilled the regularities in the SSN assignment process found in Acquisti and Gross (2009), and 2) linked the time of SSN application to two pieces of information easy to infer for most US residents (their dates and states of birth), and that the combination of process 1) and 2) rendered, in fact, SSNs predictable from demographic data.

3.2 Data and Empirical Approach

In order to test our hypothesis, we employed a fixed-effect regression model on a panel of data comprising 50 states across the 1973-2003 period. We regressed yearly, state-level rates of SSN prediction accuracy over a set of factors that may arguably impact that accuracy: a) information about the time of enactment of the aforementioned initiatives (the Interest and Dividend Tax Compliance Act, the Tax Reform Act, the Family Support Act, and the EAB), and b) demographic data, such as birth and immigration rates, by state, over time. Rates of accuracy of SSN predictions were calculated based on data from Acquisti and Gross (2009) as the ratio of complete SSNs assigned in a given state to individuals born in a given year that could be identified with fewer than 1,000 attempts using Acquisti and Gross' algorithm.

To identify the effect of tax initiatives and the EAB, we employed dummies representing the time when those policies were enacted. Our focus on state data is justified by the observation that, while SSNs are issued centrally from the SSA Baltimore's headquarters, they are initially processed at, and then assigned to, the state level.[5] The three relevant federal legislations (the Interest and Dividend Tax Compliance Act, the Tax Reform Act, and the Family Support Act) were enacted simultaneously across all states, and their dates of enactment are public knowledge. The EAB program, instead, started at different times in different states - a fact that our fixed-effect model can exploit for identification. In order to find the dates of EAB enactment by state, we combined data from a FOIA request responded by the Social Security Administration in December

[5] Specific Area Numbers are also assigned to certain US territories, which we ignore for our analysis.

2010, with data from SSA's public documents (in particular, Long 1993, SSA 1997, Donnelly 1999, Office of the Inspector General, SSA 2006, SSA 2001) and from our own email and phone interviews of SSA field offices and hospitals across the United States.

We used data from the US Census to identify the effect of birth rates. The rationale for including birth data in our regressions is that Acquisti and Gross (2009) found SSN predictions to be particularly accurate for individuals born in smaller states. We hypothesized that this was due to, primarily, different birth rates determining different numbers of SSNs applications in those states and years, and therefore different speeds of transition through the assignment scheme (slower transition implies more detectable patterns and, therefore, higher prediction accuracies). We used state-level yearly birth rates to capture this effect.

Similarly, certain categories of immigrants are allowed to receive Social Security cards and SSNs. Since the number of aliens entering a state can be quite sizable compared to the number of newborns in that state, and more populated states such as California and Florida tend to also receive more aliens *relatively* to their own newborn population (U.S. Census Bureau 2007), we included the number of legal immigrants admitted into a state in our regressions. We used the US Department of Justice Immigration and Naturalization Service's records to collect state level yearly data about *legal* immigrants admitted to the United States as a proxy for the number of those immigrants who ended up actually applying for and receiving an SSN in the same state of admission. The *direction* of the impact of aliens on the prediction accuracy for US-born individuals, however, is hard to determine on theoretical grounds. On the one hand, states and years with higher number of immigrants may exhibit less regularity in the prediction of SSNs by date of birth if their applications are concentrated and clustered in ways that add noise to the relationship between SSNs issued to US-born individuals and their dates of birth. On the other hand, if aliens' applications are spread over time, their impact on the prediction accuracy of US-born individuals may be negligible.

3.3 Regression Model

Our model is specified in the following equation:

$$Accuracy_{s,y} = \beta_0 + \beta_1 Births_{s,y} + \beta_2 Immigrants_{s,y} + \beta_3 EAB_{s,y} + \beta_4 TaxAct_y + \beta_5 FamilyAct_y +$$

$$\beta_6 DividendAct_y + \theta_y + \lambda_s + \epsilon_{s,y}$$

where *s* indexes the 50 states and *y* indexes years from 1973 (when the SSN assignment was centralized to the Baltimore SSA headquarters) to 2003 (a year for which SSN data available from the Death Master File data permits statistically robust inferences). Θ_y and λ_s are time and state fixed-effects, and $\varepsilon_{s,y}$ is the error term.

$Accuracy_{s,y}$ is the ratio (between 0 and 1) of SSNs in the Death Master File assigned in state *s* to individuals born in year *y* that could be matched with fewer than 1,000 attempts, based on data from Acquisti and Gross (2009).

$Births_{s,y}$ is the logarithm of the absolute number of births in year *y* in state *s*, based on US Census data. We expect this number to be negatively correlated with the prediction accuracy.

$Immigrants_{s,y}$ is the logarithm of the absolute number of aliens admitted into the United States in year *y* in state *s*, based on United States Department of Justice data. We do not have, *ex ante*, strong theoretical claims to predict a specific sign for the impact of this regressor on the prediction accuracy.

$TaxAct_y$, $FamilyAct_y$, and $DividendAct_y$ are dummy variables, coded as 1 after the Tax, Family, and Dividend Act had been respectively enacted, and zero before. These dummies capture the effect on the predictability of SSNs of federal initiatives that increased parents' incentives to apply for their children's SSNs at birth. We expect the coefficients for these dummies to be positive and significant.

$EAB_{s,y}$ is a dummy coded as 1 for any year y during which the EAB program was active in state s, and zero otherwise. As noted above, the date of EAB enactment for each state was established based on a combination and cross-validation of FOIA data, SSA documents, and interviews.

As noted, one feature of the EAB that makes it particularly amenable to a panel regression is that it started in different years in different states. One drawback, however, is that hospitals' participation into the EAB program is voluntary (GAO 2005) and that, at the program inception, the SSA only provided "limited education and outreach to hospitals to ensure they consistently provide information to parents regarding the timeframes for processing EAB requests" (GAO 2005, p.3). This implies that it is not always possible to unambiguously determine the date in which every, or the majority of, hospitals in a state joined the program. In order to take into account this uncertainty, we tested various versions of the model. One assumes that the EAB program started in the state in the same year when the EAB was enacted in that state; another assumes that the program was not fully implemented until the following January. Furthermore, to test the robustness of our results, we ran variations of our basic model in which we either used alternative dates for the EAB start in a given state (whenever FOIA data and public SSA documents produced slightly different start dates), or in which we outright eliminated from the regressions states for which an unambiguous EAB start date could not be established. The results we present in the following section are robust to all these modifications. Finally, since it is conceivable that the effect of EAB increased over time (as larger proportions of newborns received their SSNs under the program), we also tested a variant of the model that includes a lagged dummy representing 1, 2, or 3 years following the enactment of the EAB in a given state. Our results remain consistent also under this specification.

Time fixed effect models are often employed in the literature to estimate the impact of policy initiatives (Bertrand et al. 2004). State fixed effects allow us to control for unobserved, state-specific factors, while time dummies allow us to control for time trends. Our regressions were estimated with heteroskedastic robust standard errors clustered-corrected by state. Thus the unbiased effect of EAB can be identified from variation across state and time.

3.4 Results

The results of our regression are presented in Table 1. To avoid clutter, we do not report the coefficients for the period dummies. We present two regressions: without (first column), and with (second column) a 1-year lagged EAB dummy. Both specifications use, as start date for the actual implementation of EAB program, the January following the date the state signed in the program (the results presented here are robust to the alternative specifications of the model described above).

Results in both columns strongly support the hypothesis that the 1980s Tax initiatives and the EAB program significantly increased the predictability of Social Security numbers. Our model on average explains more than 60% of the variability in SSN prediction accuracy rates across years and states. $Births_{st}$ is significant and negative: the larger the number of births in a state, the lower the prediction accuracy. $Immigrants_{st}$ is not significant. $FamilyAct_t$ and $DividendAct_t$ are significant and positive, as hypothesized ($TaxAct_t$ is not). EAB_{st} is positive and significant, and has the largest coefficient. On average, the enactment of the EAB program and the Tax initiatives collectively add 34 percentage points to the accuracy rate of SSN predictions across all states.

These results strongly support the hypothesis that anti-fraud policy initiatives, by increasing the probability that SSNs would be applied for at birth, dramatically increased their predictability from public data. In 1972, the mean prediction accuracy for whole SSNs with fewer than 1,000 was close to 0% for most states - a level that would make brute force attacks unfeasible (see Section 4.4). By 2003, the EAB and the tax initiatives had pushed the accuracy above 50% for numerous states - which, as we discuss further below in the manuscript, exposes those SSNs to brute force attacks.

It should be noted that two additional initiatives - also in part aimed at combating identity fraud - actually contributed to make SSNs predictable from public data, and therefore vulnerable: the publication of the Death Master File, and the publication of the so-called “High Group” list. The Death Master File (which Acquisti and Gross (2009) used to predict SSNs) was made publicly available in 1981 under a Freedom of Information Act request (Wessmiller 2002, SSA 2008). Two of the goals of making it public were - paradoxically - helping identity verification and helping identity fraud prevention: in principle, DMF data could be used to expose impostors who assume deceased individuals’ SSNs. Unfortunately, the DMF has been also abused by individuals exploiting it for the very form of attack it was designed to prevent: adopting deceased individuals’ SSNs (Poulsen 2008, Hardy 2008). As for the “High Group” list (the list of the highest Group Numbers assigned to SSNs issued in a given point in time in each state), the SSA used to publish it for anti-fraud purposes. This information was meant to be used by employers to establish whether an SSN is valid. However, the High Group list can also be exploited to better infer the SSN assignment patterns and predict the assigned digits.

4 DOES THE PREDICTABILITY OF SSNs BEGET A PRACTICAL RISK OF IDENTITY THEFT?

While our regressions suggest that the predictability of Social Security numbers was significantly increased by identity fraud prevention policies, they do not tell us whether such predictability begets practical risks of identity theft for US consumers, as opposed to merely theoretical ones.

Extrapolating from the accuracy of their prediction on DMF data, Acquisti and Gross (2009) conjectured that the first five SSN digits for roughly 30 million individuals born in the US between 1973 and 2003 could be accurately matched with a single attempt, and the complete SSNs of roughly six million individuals could be identified with fewer than 1,000 attempts. These estimates do not include all SSNs issued to individuals born after 2003.[6]

Such extrapolations, however, depend on several assumptions. First, that the test results presented in Acquisti and Gross (2009) for deceased individuals' SSNs (included in the Death Master File) also hold for more realistic tests involving real-life data and Social Security numbers of alive individuals. Second, that the demographic data needed for the predictions (individuals' dates and states of birth) can be easily obtained for a vast number of US residents. Third, that it is possible to transform statistical predictions of ranges of values wherein an SSN is likely to fall into actual matches of exact SSNs, and then exploit those matches for identity theft. Fourth, that identity theft based on brute-force SSN predictions can be economically efficient for the attacker.

In this section we scrutinize those assumptions by assessing the online availability of demographic information and the availability of information systems that may be exploited for brute force, SSN prediction attacks. We begin by presenting the results of an experiment aimed at testing SSN predictions for alive individuals based on the information they publicly provided on a popular social networking site (Section 4.1). Then, we discuss the online availability of demographic data for US residents (Section 4.2), and examine both vulnerabilities and defenses in current policies, technologies, and regulations of the US online identity infrastructure (Section 4.3). Finally, we present the results of a simulation that combines different variables of interest to estimate under which conditions the predictability of SSNs may or may not raise concrete risks of widespread identity theft (Section 4.4).

4.1 Predicting Alive Individuals' SSNs

In this section we present the results of an experiment aimed at predicting the Social Security numbers of alive individuals based on the information they made available about themselves on a social networking site. Our goal in running the experiment was to assess the feasibility of real-life predictions of Social Security numbers.

4.1.1 Approach

To test whether the results presented in Acquisti and Gross (2009) would also apply to alive individuals, we mined data from the public profiles of students at a North American university who posted their names, dates of birth, and hometown information on a popular online social network (referred to below as "OSN sample"). We focused on subjects who listed dates of birth ranging from 1986 to 1990 and reported US hometowns (and therefore were likely to be US-born). Of these 621 subjects, more than 80% were born between 1986 and 1988 (therefore *before* the nationwide onset of the EAB program),[7] and most reported hometowns located in some of the more populous states, where the volume of SSNs assigned per day is higher, and therefore issuance patterns are harder to detect (Pennsylvania, New York, and New Jersey hometowns alone

[6] Being a minor is no protection against identity theft: in 2005, 5% of all identity theft reports affected victims under 18 (FTC 2006a). Furthermore, the estimates do not take into consideration the fact, as more individuals receive their SSNs under EAB, and more populate the DMF, the accuracy of predictions is likely to keep increasing over time.

represented around 50% of the sample). In other words, our sample represented a conservative test of SSN prediction accuracies, in that it corresponded to years and states where Acquisti and Gross (2009)'s prediction accuracies are worse relative to the mean prediction accuracies over the 1989-2003 period.

We hypothesized that the birth date reported in the data we mined would be correlated with the student's date of SSN application, and that the reported hometown would likely be located, for most subjects, in the same state where the actual birth occurred. Therefore, we inferred from it the *presumptive* state of SSN application. For each subject, we calculated a window of days over which to run our prediction algorithms. The window depended on the subject's presumptive state and day of SSN application.[8]

Next, we detected the most frequent first five digits assigned to *all* DMF records which were a) issued in the same state as the student's presumptive state of birth, and b) with dates of birth contained within the window of days (as calculated above) centered around the subject's self-reported date of birth. This became our one-shot prediction of the subject's first five SSN digits (his or her Area and Group Numbers). Finally, we calculated the coefficients of a (robust) OLS regression of the last four digits assigned to DMF records over their reported dates of birth, in order to infer the range of values likely to include the subject's Serial Number. Specifically, we included in the calculation *all* the DMF records which were a) issued in the same state as the subject's state of birth, and b) with reported dates of birth contained within the window of days centered around the subject's date of birth, and c) associated with the same first five digits as those predicted in the previous step. We predicted the subject's last four SSN digits by combining the

[7] Of those subjects with post-EAB birth dates, 109 were born in 1989, and only 3 in 1990.

[8] We used National Center for Health Statistics (NCHS) data to estimate the window as the number of days during a given year and in a given state it would take, on average, to assign 9,999 SSNs. It takes 9,999 assigned SSNs to transition from one Area/Group number combination (that is, the first five digits) to the following one. Since the number of SSN applications are function of births, which in both absolute and relative terms can vary significantly over time and across states, the speed of transition across SSN digits - which

subject's date of birth with the regression coefficients calculated above using DMF data (specifically, by multiplying the subject's date of birth times the regression coefficient, and adding the intercept).

To verify the accuracy of our predictions, we compared the predicted SSNs to the college's enrollment data. We employed an IRB-approved secure protocol, which had been designed *ad hoc* to disclose to us de-identified aggregate statistics about the accuracy of our predictions without disclosing to us any individual's SSN. In order to achieve that goal, we securely passed both our estimates and analysis scripts to the enrollment services at the university in question, matching individuals by their email account IDs on the university's computer system. The analysis scripts ran on the enrollment services' computers, and produced aggregate summary statistics for the accuracy of our predictions. Only these aggregate statistics were passed back to us. We focused on two accuracy metrics: whether we could correctly match the first five digits of a subject's SSN with a single attempt, and how close we could get to the last four digits, conditionally on having correctly predicted the first five.[9]

4.1.2 Results

Table 2 presents our main results. Our overall Area Number prediction accuracy (the proportion of students' first three SSN digits we correctly predicted with one attempt) was 8.5%; our Group Number prediction accuracy was 29.1%; the combined Area and Group Number accuracy (the percentage of students' first five digits we correctly predicted with a single attempt) was 6.3%. This overall prediction accuracy is close to the weighted mean accuracy of predictions (11.21%) of

depends on the volume of SSN applications over any period of time - also changes across states and years. This approach was also used in Acquisti and Gross (2009) for the prediction of DMF records' SSNs.

[9] Namely, we calculated the Euclidean distance between predicted and actual Serial Numbers. This measure is correlated with the number of attempts that an attacker who used a brute force algorithm would have to try out before hitting the correct complete SSN: for each target, we assumed that the attacker would combine the predicted Area and Group Numbers with the predicted Serial Number before moving up and down the Serial Numbers in 1-integer steps for the following attempts (keeping the predicted Area and Group Numbers constant). Different prediction algorithms are possible.

deceased individuals' SSNs (based on DMF data) reported in Acquisti and Gross (2009) for years and states corresponding to our students sample (using as weights the ratios of students from different states and different years of birth).[10]

For 33.3% of the students whose first five digits we predicted correctly, our predictions fell within fewer than 1,000 integers from the target, with the closest prediction being 191 integers away from the target's SSN. For specific post-EAB years, the prediction accuracies were much higher: for instance, 16.7% of our predicted estimates fell within fewer than 1,000 integers from the actual 9-digit SSNs for students in our sample who reported a Massachusetts hometown and a 1989 date of birth. We note that the likelihood of correctly matching a complete SSN by random guess with an error smaller than 1,000 integers would be approximately 0.0002%.

In Table 2 we also contrast the prediction accuracies in our experiment ("OSN experiment") against the appropriately weighted average predictions presented in Acquisti and Gross (2009) ("DMF experiment"). We also highlight results for the state containing the relative majority of reported hometowns in our sample (Pennsylvania) before (1986) and *during* (1989) the onset of the national EAB program. We further report the theoretical odds of accurately predicting a subject's first five digits with one attempt under the "status quo" understanding of the SSN assignment scheme.

Table 2 demonstrates that prediction accuracies on alive individuals are in the same order of magnitude as those for deceased individuals. It also confirms that both classes of predictions are hundred to thousands of times higher than what one would expect from random guess approaches.

[10] Based on natality data, we estimate that slightly more than 18 million individuals were born between 1986 and 1990 in the states represented in our OSN sample. Of them, 85,204 were dead by 2007 and their deaths were reported to the SSA - implying that their SSNs were included in the DMF records which informed our predictions. For individuals born over those years and states, the DMF dataset is sufficiently large to provide a satisfactory sample of the overall corresponding US population. Choosing a 99% confidence level, the relative standard error (RSE) of the DMF prediction accuracy is 0.97%, implying that the lower and upper bounds of the proportion of subjects in the actual population of 18 millions whose first five SSN digits are predictable with one attempt lies between 10.72% and 11.27% with 99% confidence. The smaller sample size for the OSN dataset (621 subjects) implies a larger, but still acceptable, RSE (11.42%) for the underlying proportion.

The DMF prediction accuracy (11.21%) outperforms the OSN prediction accuracy (6.3%),[11] due to potentially erroneous hometowns and dates of birth reported on the social network profiles we used in the experiment.[12] The impact of the EAB can be readily observed: for the state representing the relative majority of subjects in our sample (Pennsylvania), we contrasted prediction accuracy before (1986) and just at the onset (1989) of the EAB program. The mean ANGN prediction accuracy for Pennsylvania subjects and 1986 dates of birth was 0% (the Group Number accuracy was quite high, but all our predicted Area Numbers failed by a few integers). However, for those reporting Pennsylvania hometowns but 1989 dates of birth, the Area and Group Number prediction accuracy rose to 16.7% (Pennsylvania joined the EAB program in 1989). Similar results were also found for other states - such as Massachusetts - although their small size in the sample render such results only suggestive. The difference between the prediction accuracies for SSNs in the DMF sample issued in Pennsylvania to individuals with reported 1986 birthdates and those with reported 1989 birthdates is statistically significant at $p < 0.0001$ (Pearson $\chi^2(1) = 173.1878$, Pr = 0.000. Fisher's exact: 0.000). The same difference in our OSN experiment is statistically significant at the 5% level (Fisher's exact: 0.031).

Since the sample sizes for the PA 1986 and PA 1989 groups were relatively small, we complemented our OSN experiment with one final "natural" experiment. We focused on the subset of DMF records for individuals born in Pennsylvania in 1986 and in 1989 who *died*, specifically, in 2007. Arguably, that sample constitutes a random sub-set of all individuals who received their

[11] Assuming an underlying 11% proportion of subjects in the target population whose first five SSN digits were in fact predictable with one attempt, the OSN sample size should predict that proportion with lower and upper bounds of 7.76%-14.23% with 99% confidence. We further note that the prevalence of pre-EAB birthdates and large states in our sample created a conservative, "worse-case" test for our prediction accuracies: even in the DMF test presented in Acquisti and Gross (2009), prediction accuracies dramatically improve *after* 1989, especially for smaller states.

[12] The DMF and the OSN samples differ in important ways: while states of SSN assignment for DMF records are correct by default, and dates of birth are accurate (barring data errors), hometowns and dates of birth reported on online profiles may be *unrelated* to the time of SSN application (for instance, for individuals who received their SSN outside the EAB program), as well as *incorrect* (for instance, the individual may report as hometown a location in a different state than the state where he or she was actually born, or may intentionally provide a misleading date of birth). Both cases penalize our prediction accuracies.

SSNs in 1986 or in 1989, while also being a group as close as possible in age to our 2008 OSN sample of alive individuals. We confirmed the prediction accuracies also under this additional test: for individuals born in 1986, the Area and Group Numbers prediction accuracy was in fact 8%, and for those born in 1989 the prediction accuracy was 33%. Our results therefore confirm that predictions of *alive* individuals' SSNs are possible, and that the prediction accuracy does increase following the enactment of the EAB.[13]

4.2 Availability of birth data online

In the experiment presented in the previous section, individuals' birth data were inferred from their profiles on a popular social networking site. Since SSN predictions must be based on knowledge of an individual's state and date of birth, the availability of US residents' birth data is a precondition for prediction-based identity theft. In this section we assess the feasibility of obtaining dates and states of birth for large numbers of US residents across a variety of information services, and compare that to the costs of obtaining Social Security number in black markets.

A first source of demographic information is represented by voter registration lists. In numerous states, voters lists are public by law, and include names and dates of birth of voters.[14] For most states, this information can be obtained either directly from the respective Secretaries of State for prices ranging from $0 (New York) to $12,500 (Wisconsin), or from data brokers for prices such as 2.5 cents per record.[15]

[13] Compared to the target populations of all individuals born in Pennsylvania in 1986 and in 1989, the relative standard errors (RSE) of the prediction accuracies based on OSN data are larger (56.51% and 30.98% respectively, using 99% confidence intervals and assuming underlying true prediction accuracies of 8% in 1986 and 38% in 1989) than the RSEs of the predictions based on DMF data, due to the formers' lower sample size. Furthermore, while records of individuals born in Pennsylvania in a given year and included in the DMF can be considered a random sample of individuals born in Pennsylvania in that year, the OSN sample consists of students at a specific PA university, and is not a random sample of the PA-born population at large. However, Table 2 confirms that even the prediction accuracies in the OSN samples for PA 1986 and 1989 remain orders of magnitude larger than the prediction accuracies one would expect from informed random guesses.

[14] For instance, in the State of Washington, voters' personal data are public information under RCW 29A.08.710.

[15] See, e.g., voterlistsonline.com and http://www.whovoted.net/research.php.

Online data aggregators offer access to individual birth dates and/or personal records for negligible bulk prices (such as $20 to $30 for a month subscription which allows unlimited searches) or, in some cases, for free. These services cover hundred of millions of adult Americans.[16] For instance, usa-people-search.com claims: "[o]ur data is comprised of thousands of public records databases from across the country. We have literally billions of records, all instantly accessible. We compile data from white pages, county records, property records, and MANY other sources" (at http://www.usa-people-search.com/terms.aspx). We subscribed to two such services which advertised unlimited searches, and employed a number of back-of-the-envelope tests to gauge their coverage and accuracy of birth data, which we further describe in the Appendix. We found that these services appear to cover a significant portion of the US adult population and that they provide consistent, non-bogus data.

Infomediaries such as ChoicePoint and credit reporting agencies such as Experian, TransUnion, and Equifax continue to sell personal records (minus the SSNs) to small business and individuals on mass scale at bulk prices.

Free online people search services, without explicitly publishing individual dates of birth, make those inferrable via their extended search features.[17] We provide an exemplary test of such features in the Appendix. These services usually provide fewer personal details than social networks profiles, and are less likely to cover younger demographics. However, these services are free, and also provide phone numbers and home addresses which can be used as auxiliary information in identity theft attempts.

Other sources of demographic information include free online birthday databases;[18] numerous websites about athletes (both professional and college athletes), public figures, and celebrities;

[16] See, e.g., http://www.spokeo.com, http://www.usa-people-search.com, and http://www.peoplefinders.com.
[17] See, e.g., http://www.zabasearch.com or http://www.usa-people-search.com.
[18] Such as http://www.birthdatabase.com/query.php or http://www.stevemorse.org/birthday/birthday2.html.

commercial mailing lists aggregators (such as NextMark, http://lists.nextmark.com/); as well as mining information from public documents such as résumés or birth certificates (Griffith and Jakobsson 2005) (a reasonable but costly strategy for large amounts of records, if the records are distributed across numerous, heterogeneous sources).

As for online social networks, while not all social networking sites allow their members to publish last names and dates of birth, many profiles on the most popular sites implicitly or explicitly expose names, hometowns, and dates of birth of their members - as well as "auxiliary" information (such as phone numbers, home addresses, or pet names) which may be requested during verification processes at financial (or other) institutions. For instance, at the time of writing, Facebook counts more than 100 million members in the US alone, and has attempted to ensure that members register with their actual first and last names (O'Neill 2010). Previous studies and reports have shown that, with limited investments, third parties can automate the access and retrieval of partial or full information from large numbers of profiles in networks such as Facebook and MySpace (Gross and Acquisti 2005, Stutzman 2006, Marks 2006, Leyden 2008b, 2008a, Emery 2010). A precise estimation of US residents' birth and hometown data available through online social networks is generally not possible, because the relevant parameters continuously change.[19] To get a crude point estimate for this manuscript, in October 2010 we analyzed more than 23 thousand profiles of members of a Facebook ".edu" network (the members were associated with a North-American college institution). Of them, 46% provided a hometown, and 28% provided a date of birth.

[19] Namely: the number of members in a given network; their provision – or lack thereof – of accurate demographic data; whether they are US born, US residents, or neither; their privacy and visibility settings, which determine which information will be available to whom; the network's terms and conditions for third party developers regarding permissible usage of members' data; the network's ability to detect and interrupt mining activities from unauthorized parties; the ability of third parties to access only public, or also private, members' information; and so forth.

In short, our analysis of online people searchers, data aggregators, online social networks, and voter registration lists suggests that it is possible to obtain birth data for massive numbers of US residents at relatively low prices.

On the other hand, SSNs are, on average, costlier to obtain in mass amounts - even though the actual prices at which stolen credentials are traded in cyber-crime communities is hotly debated. Prices for stolen credentials do fluctuate over time, and black-market values are inherently difficult to scientifically assess (see, for a recent overview, Shulman (2010)). In the past, estimates of the price of SSNs in "grey" (i.e., not obviously criminal) online markets have ranged from $35 to $45 (Krim 2005). According to (ID Analytics 2006), stolen US identities (which often include SSNs) can be purchased in the black market for prices ranging from $30 to $50, although estimates of the value of SSNs in underground markets include some as low as $0.90, given the relative illiquidity of these markets (see (McCarty 2003, Thomas and Martin 2006, Herley and Florêncio 2009)). While by no means scarce, in recent years SSNs may have become harder to purchase for unauthorized third parties, as online data brokers have curtailed the sale of SSNs (GAO 2006b). Indirect evidence of the value of stolen SSNs is also offered by front companies set up by criminals to pose as legitimate businesses and "legally" purchase sensitive information from infomediaries and credit reporting agencies. In one of the most publicized data breaches in the United States, ChoicePoint was conned in 2005 by a criminal organization that purchased individual reports containing SSNs at prices "between $5 and $17" (O'Harrow Jr 2005).

4.3 Online Verification Channels

Social Security numbers are predictable in a statistical sense: the first five digits of 66% of all SSNs issued between 1988 and 2003 can be exactly predicted with just two attempts, but for the last four the algorithms presented in Acquisti and Gross (2009) only provide *intervals* of possible values wherein actual SSNs are likely to lie. This implies that many attempts are, in general, necessary to match a target's complete 9-digit SSN. Before attempting actual ID theft, an attacker therefore would need to exploit other information systems as "verification channels," testing variations of the number predicted by the algorithms, until the correct SSN digits are matched. Are there any such verification channels available?

In recent years, cybercriminal networks have started engaging in computational, brute-force attacks driven by the availability of inexpensive botnets of compromised hosts (Cooke et al. 2005). In the case of Social Security numbers, attackers may use botnets to exploit online verification channels as oracle machines (Papadimitriou 1994), in order to find which SSN corresponds to an individual with a given birth date: the attacker may repeatedly query the oracle applying an approach similar to a cryptographic dictionary attack. In the course of our research, we found a variety of online services that may be exploited as verification channels, due to the underlying conflict between applications that use SSNs as identifiers and those that use them for authentication. All of these verification channels implement defenses against attempts of abuse, but some may be vulnerable to distributed, computational attacks, given the lack of any centralized system that oversees all electronic transactions involving SSNs. In this section we only provide a brief overview of three possible verification channels to exemplify potential risks associated with an identity verification infrastructure that relies on SSNs as both individual identifiers and authenticators. We emphasize that the verification strategies described below are informed inferences based on publicly available data about those channels and the attack strategies that criminals are known to be using.

- The SSN Verification Service. One potential vector for verification attacks is maintained by the SSA itself: the SSN Verification Service (SSNVS) allows companies and self-employed individuals to verify online other individuals' SSNs based on their names and dates of birth

(SSA 2007c). Usage of the service is conditional to previous registration and - therefore - security vetting by the SSA. Such vetting process *may* be bypassed or fouled *if* an attacker succeeded in compromising the credentials of already vetted accounts.

- The E-Verify System. The US Citizenship and Immigration Services's (USCIS) E-Verify system allows the online verification of SSNs. An attacker who succeeded in impersonating an employer (or a set of employers), or who could steal legitimate employers' password, may therefore attempt to abuse the system from multiple accounts for SSN verification purposes.[20]
- Instant Credit Approval Services. In the private sector, numerous online "instant credit card application" services (or similar wireless carriers and "instant" lending companies verification services (ID Analytics 2005)) take as input an individual's name, date of birth, and SSN (plus, sometimes, auxiliary information such as his address or phone number) and return as output a preliminary approval or denial of that individual's application for credit or service.

Significantly more details about these and other verification channels (from "spear" phishing attacks and "synthetic" identity theft, to the mining of public records that redact the first five SSN digits but still include an individual's last four digits of the SSN as identifiers) are reported in the Appendix, including a discussion of potential defenses against attacks.[21]

4.4 Are Brute Force SSN Predictions Economically Effective?

To complete our discussion of whether the predictability of SSNs begets practical or merely theoretical risks of identity theft, we employ a computational model of an attacker's costs and benefits of carrying out identity theft by brute force SSN predictions. We combine into the model

[20] As of March 2011, the Department of Homeland Security launched a "self-check" program to US residents to use E-Verify in order to test their own eligibility for employment before looking for jobs. See http://edition.cnn.com/2011/US/03/21/worker.eligibility/.

[21] Synthetic identity theft does not rely on knowledge of a victim's date of birth, but merely the correct correspondence between a certain SSN and the birth data and state of the individual to whom that SSN was issued. In such attacks, criminals combine combine *fake* names and addresses with a *real* individual's SSN and the correct dates of birth associated with that SSN (Hoofnagle 2007a).

data about the accuracy of alive individuals' SSN predictions, estimates of the availability of birth data for US residents, and projections about the feasibility of exploiting online verification channels for SSN matching. By simulating different scenarios based on possible values of the parameters, we assess under what conditions the predictability of SSNs provides represents an economically effective or ineffective vector for identity theft - and therefore whether such predictability actually exposes US residents to concrete, or only theoretical, risks of identity theft.

We consider a model in which the attacker obtains states and dates of birth for US residents through one of the vectors described in Section 4.2; predicts the target's SSN using on the algorithms described in Acquisti and Gross (2009); and then rents a botnet to apply for credit cards via online instant credit card approval services, cycling through a range of values around each predicted SSN, as described in Section 4.3. The model is populated with demographic data from the US Census, SSN prediction accuracy rates from Acquisti and Gross (2009), and ranges of costs of rented botnets and prices of stolen credentials obtained from academic and industry reports. Our goal is to estimate how the economic effectiveness of identity theft attacks based on the prediction of SSNs varies as function of the parameters values.

4.4.1 Approach

We consider an attacker that targets individuals born in state s in year y. The attacker rents a botnet with nIp IP addresses, and uses it to connect to online instant credit card approval services, applies for credit cards using the victim's name, DOB, and predicted SSN, and then cycles through variations of predicted SSNs. We calculate the number of potential targets born in a given state and year from the number of births $nB_{s,y}$ (based on vital statistics: e.g., NCHS (1998)). We use the number of attempts $nAtt_{s,y}$ required, on average, to determine an SSN with success rate s_p (from (Acquisti and Gross 2009)) to calculate the number of "credentials" (that is, matched SSNs) the attacker will be able to produce. We assume that an IP address gets blacklisted by an online credit card issuer after *k* incorrect credit card applications, or "attempts;" that the criminal distributes his attacks across *nCra* credit card services, or "issuers"; that he can find birth data for a ratio b_p of the potential targets; and that 7 out 9 digits are sufficient for a CRA to answer an issuer's inquiry with a positive match in c_p percent of the cases (ID Analytics 2006). Since the attacker wants to avoid detection by the CRAs that process applications forwarded by the credit approval sites, we set a hard-coded attacker's constraint that self-limits traffic to t_p percent of the daily volume of CRA inquiries (estimated at 4M per day by the FTC in 2004; see FTC 2004). We consider a marginal cost of *c* per day to rent a botnet with a thousand IP addresses. Finally, we define as *p* the variable price at which each correctly determined credential can thereafter be sold in the black market.

4.4.2 Model

Using these definitions, we compute the number of potential targets for a given state *s* and year *y*,

$$nTargets_{s,y}=nB_{s,y}*b_p \qquad (1)$$

as percentage of the population born in that state and year for which birth information is available. Using the number of attempts necessary for the Acquisti-Gross algorithm to correctly predict an SSN we obtain:

$$nReqatt_{s,y}=nTargets_{s,y}*nAtt_{s,y} \qquad (2)$$

where $nReqatt_{s,y}$ is the number of required attempts for targets born in state *s* in year *y*. The number of attempts *available* to the attacker follows from our assumptions about the botnet and CRAs:

$$nAvatt=nIp*nCra*k \qquad (3)$$

Finally, we compute the number of attempts that can be made per CRA per minute as

$$nReqMin{=}t_p{*}(4000000/1440) \qquad (4)$$

We estimate the number of computed credentials $nCred_{s,y}$ as

$$nCred_{s,y}{=}\begin{cases} nTargets_{s;y}{*}s_p{*}c_p \text{ if } nReqatt_{s;y}{\leq}nAvatt \\ \frac{\mathrm{nAvatt}}{\mathrm{nAtt}_{\mathrm{s;y}}}{*}\mathrm{s_p}{*}\mathrm{c_p} \qquad \text{otherwise} \end{cases} \qquad (5)$$

The credentials can be obtained in

$$nMinutes_{s,y}{=}\begin{cases} \frac{nReqatt_{s;y}}{nReqMin{*}nCra} \text{ if } nReqatt_{s;y}{\leq}nAvatt \\ \frac{\mathrm{nAvatt}}{\mathrm{nReqMin * nCra}} \qquad \text{otherwise} \end{cases} \qquad (6)$$

The cost of renting a botnet (with billing in hours) is

$$cTot_{s,y}{=}\frac{\frac{nIp}{1000}{*}c}{24}{*}\frac{nMinutes_{s,y}}{60} \qquad (7)$$

An identity theft operation based on brute force SSN predictions would therefore generates a profit or loss (gross of the fixed costs of obtaining birth data and writing the scripts) of:

$$profit{=}nCred_{s,y}{*}p{-}cTot_{s,y} \qquad (8)$$

4.4.3 Results

Table 3 summarizes different values of the parameters we chose for our simulation. We assume that there exists 20 independent online instant credit approval services (a conservative assumption) that a third party can attempt to exploit to verify predicted SSNs. The attacker is renting a small botnet of 10,000 nodes/IP addresses for an entire day (even though some of the verification algorithms could be run within a few hours; the assumption account for bots' downtime). Based on estimates reported in Lesk (2007), we assume a botnet rental cost of $1,000 per day for 10,000 bots (as noted above, however, cost estimates for renting a botnet vary widely in the literature). Citing industry data, Moore et al. (2009, p.3) report that the "information needed to apply for credit in someone else's name, such as name, social security number, and birthday, fetches $1 to $15 per set." However, as noted above, estimates for the price at which a valid SSN can be sold on the black cybermarkets also range widely (Herley and Florêncio 2009), from a few cents to more than $40. Hence, we consider three different scenarios: $0.10, $12, and $25.

In Figure 1, we show profits, in thousands of dollars, earned using brute-force SSN prediction attacks and by targeting individuals born across all states from 1973 through 2003 for different credential prices. States are sorted from 1 to 50 in increasing order of number of births in 1973. A profit level of zero implies that the third party would not attempt predictions for individuals born in a certain state and year, since the expected probability of prediction success cannot guarantee a positive profit and therefore does not justify the investment.

The results imply that brute force, SSN prediction-based identity theft attacks are economically unfeasible when the prices at which criminals can sell obtained credential are extremely low ($0.10), or when the targets are born before the mid 1980s in larger states. Under such conditions, the costs involved in renting botnets for the attacks are too large compared to the expected revenues. However, the results show that brute-force attacks are, instead, almost always profitable at higher credential prices on targets born after the mid 1980, and particularly so for those borne in smaller states: under medium and high credential prices, the profit rates can be as high as a thousand percent.

To illustrate the range of possible outcomes as function of other parameters' values, we also present in the Appendix the profitability of SSN prediction-based attacks against victims born on an arbitrarily chosen post-EAB year of birth (1991). We consider there three scenarios: one using a conservative set of parameters, a liberal set of parameters, and the mean between the two.[22] We also assume a middle-of-the-road credential sale price of $12.

4.4.4 Opportunity Costs of Computational Attacks

So far we have shown that, depending on the price of stolen credentials in underground cybermarkets and on the choice of target, SSN prediction-based attacks range from economically unfeasible to highly profitable. From the attacker's perspective, whether such attacks are economically desirable depends on their opportunity costs: namely, on the relative profitability of other identity theft attack vectors such as data breaches, phishing attacks, purchasing victims' credentials in the black market, or other offline scams. In this section, we consider the different economic factors that impact such assessment. We focus on the comparison of a) overall operation costs and b) scalability of attack across alternative strategies. For instance, purchasing stolen credentials and conducting offline scams both carry high variable costs, and therefore are not highly scalable attacks; while phishing attacks are highly scalable, and attacks based on data breaches, while not scalable, can produce massive amounts of stolen credentials.

As calculated in Equation 8, profits are net of botnet rental costs, but do not account for the costs of obtaining demographic information and writing code to run brute force attacks, because the latter tend to be fixed costs and quite negligible. The tangible cost of obtaining demographic information for US residents, using the channels described in Section 4.2, varies from zero (e.g., in states where voter registration information is public), to moderate fixed fees (such as $29.99 a month to access www.usapeoplesearch.com databases, or the opportunity costs of writing scripts to

[22] For the conservative estimate, we set the percentage of potential targets that the attacker can find birth data to b_p=0.1, and the success rate with which SSNs are correctly estimated to s_p=0.5. We furthermore assume that IP addresses are blocked after k=3 incorrect attempts. For the liberal estimate we set b_p=0.5,s_p=0.9 , and k=10.

download data from other online repositories). The costs of planning and executing brute force attacks (other than the aforementioned rental of botnets) tend to be themselves opportunity costs (the time invested in developing the code to command bots, predict SSNs, and test predictions across credit card issuers). As noted in Section 4.3, computational attacks have become quite common in various forms thanks to the commoditization of tools of attacks (Anderson et al. 2008, Moore et al. 2009) and the increasing availability of cheap botnets. Furthermore, computational attacks such as brute force SSN predictions and phishing attacks can be decentralized, operated remotely (limiting the attackers' expected costs of identification and prosecution), and are highly scalable (once coded, the attacks can be automated, leading to average costs per credential obtained that decrease with the number of targets).[23] Brute force predictions of SSNs are "stealth" forms of identity theft, harder to detect than a data breach, and more customizable based on desirable socio-economic traits of the victims. Paradoxically, as investments into securing corporate databases containing sensitive personal information increase, computational attacks that bypass them may become more economically appealing.

On the other hand, attacks based on data breaches can produce massive amounts of credentials with a very focused investment. Such attacks, however, often carry high fixed costs (obtaining valuable credentials from large corporate databases tends to be time consuming, and the efforts are not necessarily transferrable to other targets: consider, e.g., the TJ Max attack: see United States Department of Justice (2008)).

[23] The vulnerability of financial services to brute-force attacks - especially those made possible by botnets - is a concern even when strong passwords and "three strikes" rules (an IP address or an account get blacklisted after three failed logins or attempts) are employed (Hole et al. 2006), since a rational attacker will ensure that attacks (such as login attempts) "are launched from diverse IP addresses" (Florêncio et al. 2007). Even "compliance [to security standards] does not guarantee security" to threats such as brute force attacks (Meyran 2008).

5 CAN CURRENT SSNs POLICIES EFFECTIVELY COMBAT IDENTITY THEFT?

We summarize what we have established so far in this manuscript in Figure 2, first row. The predictability of Social Security numbers was made possible by a series of policy initiatives originally intended to avoid fraud or identity theft. However, alone, the predictability of Social Security numbers would not constitute a threat. The threat, as Figure 2 shows, only arises from a chain of vulnerabilities: the wide availability of personal information feeds the algorithms that exploit the predictability of SSNs; the availability of compromised bots allows distributed verification attacks; our markets' reliance on credentials-based systems offers verification channels for those predictions; in turn, such predictions would not raise significant concerns if it were not for the fact that Social Security numbers, once designed for identification, are now used as authenticators in many sensitive applications. Together, this chain of vulnerabilities give rise to a systemic privacy risk, a vulnerable information ecosystem in which identity theft based on SSN predictions may not just be technologically feasible but economically viable. In the rest of this section we highlight possible defense strategies against these risks.

5.1 The Current Policy Approach

Industry and legislators interested in addressing the problem of identity theft have, so far, focused on two strategies.

A first strategy - widely adopted by financial institutions in the US - consists of extending beyond SSNs and associated names and dates of birth the number of questions one needs to answer in order to access an account. These are the so-called "out of wallet" questions, such as the name of one's pet or best friend, or the individual's high school. Increasing the number of questions consumers have to answer correctly in order to access (or modify) financial and other services also carries unintended costs: it increases the cognitive burden for consumers, therefore raising the probability of "false positive" denials of service; it requires that such organizations store the answers to the various challenge questions (making those answers vulnerable to data breaches); and relies on information which could be easy to assemble for a third party: "[o]ut-of-wallet questions have become well known to fraudsters" (ID Analytics 2003, p. 22), who can purchase answers to the questions or find them from numerous sources.[24] More importantly, increasing the number of verification questions may deflect "current account" frauds, but are often ineffective against "new account" frauds (in which knowledge of SSNs and associated dates of birth is often sufficient to create fraudulent accounts: see Section 2). It also does not correct the underlying vulnerability of a system based on using as passwords information shared by multiple parties which cannot be revoked or replaced.

A second, relatively more recent strategy we have highlighted in the Introduction focuses on limiting the public exposure of SSNs, or redacting some of their digits from public documents:

1. Approach 1: Removing the presence of SSNs in public documents and online services (GAO 2008) and limiting their trade.[25]
2. Approach 2: Redacting the first five digits of SSNs from public documents.[26]

Such policies face hurdles similar to those associated with out-of-wallet questions, and reflect a belief ostensibly shared by both the private sector and public sector entities: while SSNs are acknowledged to be too widely available, they are not considered *so* available to make them

[24] Auxiliary data often requested in financial verification systems - such as phone numbers, home addresses, pet names, or mothers' maiden names - may also be inferred online from social networking site profiles (Gross and Acquisti 2005), free people-search services, or publicly available databases (Griffith and Jakobsson 2005).

[25] For instance, President Bush put an end to the mandatory use of SSNs for federal identification with Executive Order 13478. As noted in Section 2, in the last three years a number of federal bills such as H.R. 3046 of 2007, H.R. 948 of 2007, S. 238 of 2007, S. 2915 of 2008, and S. 1618 of 2009 aimed at limiting the display or sale of SSNs, but did not become law. However, members of Congress can reintroduce bills that did not come up for debate under new numbers in the next session.

[26] See http://www.ncsl.org/programs/lis/privacy/SSN2007.htm, as well as GAO (2006b).

useless for authentication.[27] For instance, the recent policy recommendations of the President's Identity Theft Task Force focused on limiting "unnecessary" uses of SSNs in order to preserve their role as integral part of the financial system, under the presumption that SSNs can, in fact, remain private (The President's Identity Theft Task Force 2007). Similarly, the Government Accountability Office, in 2008, noted that policy initiatives in the area of SSN protection struck "an appropriate balance between protecting SSNs from misuse and making a portion available for appropriate parties to firmly establish the identity of specific individuals" (GAO 2008, p.7). The attention towards protecting SSNs is also reflected in the focus that both the SSA and other government bodies have placed on *individuals*' responsibility to keep they own SSN private and protected (see FTC 2006b, SSA 2007b, The President's Identity Theft Task Force 2007, FTC 2007, 2008).

Asking individuals to protect their SSNs, however, causes them costs without assurances of benefits: individuals have little to no control upon how their SSNs are used, collected, and traded among so-called infomediaries (such as CRAs and data aggregators); furthermore, their SSNs remain predictable from otherwise public data. The FTC (2008), in fact, recognizes that "[l]imiting the supply of SSNs that are available to criminals [...] is more complex. SSNs already are available from many sources, including public records, and it may be impossible to 'put the genie back in the bottle.'" Therefore, at the very least, the methods by which businesses authenticate new and existing customers should be strengthened beyond the usage of SSNs (FTC 2008).

Recent legislative interventions in this area, which attempt to protect SSNs in order to preserve their role as authenticators in the credit/financial infrastructure, while certainly well-meaning therefore risk of being ineffective, or even counterproductive. Protecting SSNs by making them less available (Approach 1) makes sense only until one realizes that SSNs effectively cannot be

[27] See, for instance, Irwin Financial Corporation (2007, p.3): "Are SSNs so widely available that they should never be used as an

removed from public knowledge, because they are predictable from public data. In this case, and paradoxically, the attempts to protect them by limiting their visibility may backfire, in that - by creating scarcity of SSNs - those attempts may make SSNs even more trusted as passwords, while rendering them even more valuable commodities for cybercriminals. In other words, by making them less available to legit third parties, but not less valuable to criminals, such initiatives would not curtail the criminal *demand* of SSNs but only its *supply*. This may, in turn, make it more likely that attackers engage in mass pharming of such data. Going back to the simulations we presented in Section 4.4, if the price of compromised SSNs raises, so does the profitability of engaging in computational attacks (see Figure 1, (b)).

Furthermore, because the first five digits of assigned SSNs are particularly easy to predict, any legislative initiative focused on their redaction from public documents (Approach 2) seems questionable, since those initiatives essentially expose what is harder to predict (the last four digits) and protect what it easy to identify (the first five). Similarly, many organizations' current practice of asking consumers for the last four digits of their SSNs (or using those digits as default passwords in certain systems, or listing them as identifiers in bills or other documents), also appears ill-conceived.

5.2 Alternative Strategies

We highlight a number of alternative defensive strategies in Figure 2, second row.

Consider Block A: curtailing the availability of personal information online. In principle, social networking sites may try to discourage the revelation of birth data by tuning default privacy settings, and other services from which birth information can be inferred may reassess data accessibility policies. In reality, information already available cannot be taken back, and given our

authenticator? NO. However, we agree that it is prudent to use partial SSNs whenever possible."

markets' reliance on the free flow of information, it is unlikely that birthdata for US residents can be taken back from the public domain, or made harder to be obtained by third parties. Hence, this does not seem like a viable strategy.

Similarly, consider Block C. Computer scientists are researching ways to protect individual PCs from being compromised and abused (for instance, through trusted computing solutions). However, it is unlikely that in the short term we will be able to solve the problem of botnets and computational attacks.

Consider Block B: randomizing the assignment of SSNs would offer temporary relief for newly issued SSNs. In fact, the SSA has recently announced that it will switch to a randomized assignment scheme beginning June 2011 (SSA 2011a). While praiseworthy, however, the initiative leaves hundreds of millions of Social Security numbers issued under the current scheme still vulnerable (SSA 2011b). Furthermore, it does not address the ultimate cause of the vulnerability of US identity infrastructure: the usage of SSNs for both identification and authentication purposes.

Consider Block D: credit bureaus may coordinate their efforts to share credit application information in real time, to make crimes such as synthetic identity theft harder to perpetrate. In Figure 3, we present the results of a new simulation in which the three-strikes rule is implemented in coordination by the 20 online credit card issuers, so that if an IP address is blocked by one issuer, it is also immediately blocked by others (similar strategies currently are used by merchants that accept credit card *payments*, thanks to the collaboration between companies that provide online transactions; but do not apply to credit *requests*, which - as highlighted above - take time to propagate through the system). Using the model presented in Section 4.4, we can estimate how these initiatives would disrupt the underground cybermarket. In Figure 3, we plot profits for the case in which CRAs communicate in real time, flagging an offending IP after (in this case) 3 attempts across *all* CRAs, instead of just for one. Profits would be reduced dramatically, rendering this approach much less lucrative. As suggested by Figure 3, such coordination, alone, could dramatically reduce the profitability of computational attacks, by limiting the space of verifications that attackers can try. Furthermore, although services such as SSNVS and E-Verify already employ a variety of security mechanisms, those mechanisms should be specifically audited in light of the discovery that brute force attacks with even just a few dozen or hundred repetitions on somebody's SSNs may be sufficient to identify that person's SSN.

5.3 Abandoning SSNs as Means of Identification and Authentication

None of the defense strategies we just considered is foolproof, and most may carry unintended consequences. Furthermore, most of the above strategies only help *future* SSN applicants; and none of them addresses the inherent tension between SSNs as shared identifiers and SSNs as sensitive authenticators, which ultimately is the main culprit of identity theft in the United States. Protecting US online and offline identities, therefore, may require a more aggressive reconsideration of our policies and information systems for identity verification. One such solution would consist of banning, through sunset provisions, the usage of SSNs as passwords and as authenticators. This would force private and public sector to resort to alternative forms of authentication (Figure 2, Block E). Rather than allocating resources to keeping an insecure identity infrastructure function through a patchwork of costly ex-post interventions (including identity theft insurances, out-of-wallet questions, and so forth), investments could be allocated to the deployment of usable cryptographic protocols, already well researched in the computer security community. Identity management systems based on blind signatures and anonymous credentials (see, e.g., Chaum 1983, Chaum 1985, Brands 2000, Camenisch and Lysyanskaya 2001), for instance, can address the need for data sharing as well as the need for data privacy: such protocols can be used for authentication without identification, as well as for authentication that does not allow the authenticator to later impersonate the authenticated.

In principle, such solutions have nowadays become sufficiently affordable and usable to be mass-deployed (for instance, every Internet browser already contains cryptographic capabilities - completely transparent to the user - to allow the the completion of secure online purchases). Deploying such tools on a massive scale, however, would not be costless. A careful economic analysis could highlight the short and long term cost and benefits associated with transitioning from an SSN-based to a cryptographic credential-based identity infrastructure.

6 CONCLUSIONS

In this manuscript we have analyzed the policy roots and implications of the predictability of Social Security numbers. Our results indicate that a number of well-meaning public policy initiatives – paradoxically intended to prevent identity fraud – introduced regularities in the issuance of SSNs that make them predictable (Section 3). Furthermore, they indicate that such regularities, coupled with the increasing public availability of personal data and known (but unresolved) vulnerabilities in the US identity infrastructure, make SSNs effectively predictable based entirely on public information, and therefore heightens statistical, but nevertheless concrete, risks of identity theft (Section 4). Finally, our results indicate that recent legislative initiatives in the area of identity fraud prevention, while well intended, may be misguided, as they attempt to remove SSNs from public view when SSNs are, effectively, semi-public information (Section 5). Removing them from public access may increase the public’s trust in their value as means of authentication, but would not make them less predictable. Even the recent plan to randomize the assignment of SSNs beginning June 2011 (SSA 2011a) (following the publication of Acquisti and Gross (2009)’s study), would leave a hundred of millions of already issued Social Security numbers potentially vulnerable (since older SSNs assigned under the previous scheme will *not* be replaced: see SSA 2011b). Furthermore, randomizing the issuance of new SSNs does not address the inherent vulnerability of using the same sequence of digits for both identification and authentication purposes.

Our findings do not imply that policies in this area are inherently ineffective and counterproductive. The predictability of SSNs is the unintended and unpredictable consequence of the complex interaction of multiple, ostensibly unrelated, policy initiatives. Our results, instead, highlight the importance of considering holistically the interplay of policies, technology, and human behavior in formulating strategies for the protection of personal data. If anything, our findings point out the need for more aggressive policy interventions, such as altogether prohibiting the usage of SSNs as authenticators in the private sector (Solove 2003), or in fact making SSNs entirely public (LoPucki 2003), in order to effectively terminate their usage as authenticators.

Our insecure identity infrastructure costs individuals, companies, and the government tens of billions of dollars per year when attacks are successful - but even more significant are the costs due to wasteful investments aimed at prolonging the status quo system against ever mounting cyber-threats. Protecting SSNs by removing them from public view is not a viable solution if SSNs remain predictable from public data. In place of SSNs, identity solutions based on cryptographic tools such as blind signatures and digital credentials may be more apt to resolve our information economies' dual needs of keeping personal identifiers public while sensitive authenticators private.

References

A. Acquisti and R. Gross. Predicting Social Security numbers from public data. *Proceedings of the National Academy of Science*, 196 (27): 10975–10980, 2009.

R. Anderson, R. Bohme, R. Clayton, and T. Moore. Security economics and the internal market. Report to the European Network and Information Security Agency, 2008.

M. Bertrand, E. Duflo, and S. Mullainathan. How Much Should We Trust Differences-in-Differences Estimates? *Quarterly Journal of Economics*, 119 (1): 249–275, 2004.

G. Block, G.M. Matanoski, and R.S. Seltser. A method for estimating year of birth using Social Security number. *American Journal of Epidemiology*, 118 (3): 377–395, 1983.

S.A. Brands. *Rethinking public key infrastructures and digital certificates: Building in privacy*. The MIT Press, 2000.

J. Camenisch and A. Lysyanskaya. An efficient system for non-transferable anonymous credentials with optional anonymity revocation. In *Eurocrypt*, pages 93–118, 2001.

D. Chaum. Blind signatures for untraceable payments. In *Advances in Cryptology: Proceedings of Crypto*, volume 82, pages 199–203, 1983.

D. Chaum. Security without identification: Transaction systems to make big brother obsolete. *Communications of the ACM*, 28 (10): 1030–1044, 1985. ISSN 0001-0782.

C. Clifton and D. Marks. Security and privacy implications of data mining. In *Proceedings of the 1996 ACM SIGMOD Workshop on Data Mining and Knowledge Discovery*, pages 15–19. ACM, 1996.

M. Cook. The lowdown on fraud rings. *Collections & Credit Risk*, 10, 2005.

E. Cooke, F. Jahanian, and D. Mcpherson. The zombie roundup: Understanding, detecting, and disrupting botnets. In *Workshop on Steps to Reducing Unwanted Traffic on the Internet (SRUTI)*, pages 39–44, June 2005.

G. Donnelly. Subject counterfeiting and misuse of the social security card and state and local documents, 1999. http://www.ssa.gov/legislation/testimony_072299.html.

D. Emery. Details of 100m Facebook users collected and published. *BBC News*, July 28, 2010. http://www.bbc.co.uk/news/technology-10796584.

Executive Order. Amendments to executive order 9397 relating to Federal Agency use of Social Security Numbers. *Federal Register*, 73 (225), 2008.

D. Florêncio, C. Herley, and B. Coskun. Do strong web passwords accomplish anything? In *HOTSEC'07: Proceedings of the 2nd USENIX workshop on Hot topics in security*, pages 1–6, Berkeley, CA, USA, 2007. USENIX Association.

FTC. Report to Congress under sections 318 and 319 of the Fair and Accurate Credit Transactions Act of 2003, 2004. http://www.ftc.gov/reports/facta/041209factarpt.pdf.

FTC. Identity theft complaints by victim age, 2006a. http://www.ftc.gov/sentinel/reports/Sentinel_CY-2005/idt_victim_age.pdf.

FTC. Take charge: Fighting back against identity theft, 2006b. http://www.ftc.gov/bcp/edu/pubs/consumer/idtheft/idt04.pdf.

FTC. Staff summary of comments and information received regarding the private sector's use of Social Security numbers. FTC, Division of Privacy and Identity Protection, Bureau of Consumer Protection, 2007. http://www.ftc.gov/bcp/workshops/ssn/staffsummary.pdf.

FTC. Security in numbers: SSNs and ID theft: A FTC report providing recommendations on Social Security number use in the private sector. FTC Report, December 2008, 2008. http://www.ftc.gov/os/2008/12/P075414ssnreport.pdf.

GAO. Social Security numbers: Use is widespread and protections vary. GAO-04-768T, 2004. http://www.gao.gov/new.items/d04768t.pdf.

GAO. SSA: Actions needed to strengthen processes for issuing social security numbers to children. GAO-05-115, 2005. http://www.gao.gov/new.items/d05115.pdf.

GAO. Social Security numbers: More could be done to protect SSNs. Testimony Before the Subcommittee on Social Security, Committee on Ways and Means, House of Representatives., 2006a. http://www.gao.gov/new.items/d06586t.pdf.

GAO. Internet resellers provide few full SSNs, but Congress should consider enacting standards for truncating SSNs, 2006b. http://www.gao.gov/new.items/d06495.pdf.

GAO. Social security numbers are widely available in bulk and online records, but changes to enhance security are occuring. GAO-08-1009R, 2008. http://www.gao.gov/new.items/d081009r.pdf.

A. Greenberg. ID theft: Don't take it personally. *Forbes.com*, 2010. http://www.forbes.com/2010/02/09/banks-consumers-fraud-technology-security-id-theft.html.

V. Griffith and M. Jakobsson. Messin' with Texas: Deriving mother's maiden names using public records. In *Proceedings of the Third Conference on Applied Cryptography and Network Security*, volume 3531, pages 91–103. Springer LNCS, 2005.

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*, pages 71–80. ACM, 2005.

M. Hardy. SSA lists thousands of live persons as dead. *Federal Computer Week*, June 26, 2008. http://www.fcw.com/online/news/152975-1.html.

C. Herley and D. Florêncio. Nobody sells gold for the price of silver: Dishonesty, uncertainty and the underground economy. In *Eighth Workshop on the Econimics of Information Security (WEIS 2009)*, 2009.

K.J. Hole, V. Moen, and T. Tjostheim. Case study: Online banking security. *Security & Privacy, IEEE*, 4 (2): 14–20, 2006.

C.J. Hoofnagle. Identity theft: Making the known unknowns known. *Harvard Journal of Law and Technology*, 21 (1): 98–122, 2007a.

C.J. Hoofnagle. How SSNs are used to commit ID theft: Synthetic identity theft. Presentation at the FTC Workshop on Security in Numbers: SSNs and ID Theft, December 10-11, 2007b. http://www.ftc.gov/bcp/workshops/ssn/presentations/Hoofnagle.pdf.

ID Analytics. National report on identity fraud, 2003. http://www.idanalytics.com/. Document obtained from the company.

ID Analytics. U.S. national fraud ring analysis: Understanding behavioral patterns, 2005. http://www.idanalytics.com/. Document obtained from the company.

ID Analytics. National data breach analysis, 2006. http://www.idanalytics.com/. Document obtained from the company.

Irwin Financial Corporation. Comment to FTC's SSNs in the Private Sector, Project No. P075414, 2007. http://www.ftc.gov/os/comments/ssnprivatesector/531096-00310.pdf.

J. Krim. Net aids access to sensitive ID data: Social Security numbers are widely available. *Washington Post*, April 4: A01, 2005. http://www.washingtonpost.com/wp-dyn/articles/A23686-2005Apr3.html.

M. Lesk. The new front line: Estonia under cyberassault. *IEEE Security & Privacy*, 5 (4): 76–79, 2007.

J. Leyden. Spammers open new front on social networking sites. The Register, May 14 2008, 2008a. http://www.theregister.co.uk/2008/05/14/social_network_spam/.

J. Leyden. Myspace wins record $230m judgement against spammers. The Register, May 14 2008, 2008b. http://www.theregister.co.uk/2008/05/14/myspace_spam_ruling/.

W.S. Long. Social Security numbers issued: A 20-year review. *Social Security Bulletin*, 56 (1): 83–86, 1993.

L.M. LoPucki. Did Privacy Cause Identity Theft? *Hastings Law Journal*, 54 (4), 2003.

P. Marks. Pentagon sets its sights on social networking websites, 2006.

B. McCarty. Automated identity theft. *IEEE Security and Privacy*, 1 (5): 89–92, 2003.

Ron Meyran. Financially motivated hacking: the non-vulnerability challenge. *Financial Services Technology*, 2008.

T. Moore, R. Clayton, and R. Anderson. The economics of online crime. *The Journal of Economic Perspectives*, 23 (3): 3–20, 2009.

NCHS. Births, marriages, divorces, and deaths for 1997, 1998. http://www.cdc.gov/nchs/data/mvsr/mv46_12.pdf.

E. Ody. Identity theft fraud falls 34%, victims pay more. *Bloomberg.com*, 2011. http://www.bloomberg.com/news/2011-02-08/identity-theft-fraud-falls-34-victims-pay-more.html.

Office of the Inspector General, SSA. Follow-up of the Enumeration at Birth program. Audit Report, April 2006, A-08-06-26003, 2006. http://www.ssa.gov/oig/ADOBEPDF/audittxt/A-08-06-26003.htm.

Office of the Inspector General, SSA. State and local governments' collection and use of Social Security numbers. Audit Report, September 2007, A-08-07-17086, 2007. http://www.ssa.gov/oig/ADOBEPDF/audittxt/A-08-07-17086.htm.

R. O'Harrow Jr. ID data conned from firm: Choicepoint case points to huge fraud. *Washington Post*, February 17: E01, 2005. http://www.washingtonpost.com/wp-dyn/articles/A30897-2005Feb16.html.

N. O'Neill. Verify your facebook account or wind up frustrated. *All Facebook*, September 21, 2010. http://www.allfacebook.com/verify-facebook-account-2010-09.

C. Papadimitriou. *Computational Complexity*. Addison-Wesley, 1994.

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

K. Poulsen. Feds charge california woman with stealing IDs from the dead. *Wired Blog Network: Threat Level*, April 17, 2008. http://blog.wired.com/27bstroke6/2008/04/feds-charge-cal.html.

S. Romanosky, R. Telang, and A. Acquisti. Do data breach disclosure laws reduce identity theft? *Journal of Policy Analysis and Management*, 2011. In Press.

Samuelson Law, Technology & Public Policy Clinic. Comment to FTC's SSNs in the Private Sector, Project No. P075414, 2007. http://www.ftc.gov/os/comments/ssnprivatesector/531096-00295.pdf.

A. Shulman. The value of your credentials. *Help Net Security*, October 14, 2010. http://www.net-security.org/article.php? id=1508&p=5.

R.E. Smith. *Ben Franklin's web site*. Privacy Journal, 2000.

R.E. Smith. Social Security numbers: Uses and abuses, 2002. http://www.simson.net/ref/databasenation/SSNReport2001.pdf.

D.J. Solove. Identity theft, privacy, and the architecture of vulnerability. *Hastings Law Journal*, 54: 1227–1252, 2003.

SSA. Program operations manual system (POMS). SSA Pub. No. 68-0100201, November 8, 1996. https://s044a90.ssa.gov/apps10/poms.nsf/.

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

SSA. Audit of Enumeration at Birth Program. Office of The Inspector General, SSA, A-08-00-10047, 2001. http://www.ssa.gov/oig/ADOBEPDF/audittxt/A-08-00-10047.htm.

SSA. Protecting the integrity of Social Security numbers (SSNs). *Federal Register*, 72 (127), 2007a.

SSA. Identity theft and your Social Security number. SSA Publication No. 05-10064, 2007b. http://www.ssa.gov/pubs/10064.html.

SSA. Social Security number verification service (SSNVS) Handbook, 2007c. http://www.ssa.gov/employer/ssnvs_handbk.htm.

SSA. Is SSA's death master file available online? http://ssa-custhelp.ssa.gov/cgi-bin/ssa.cfg/php/enduser/std_adp.php? p_faqid=149, 2008.

SSA. Social security number randomization, 2011a. http://www.ssa.gov/employer/randomization.html.

SSA. Social security number randomization frequently asked questions, 2011b. http://www.ssa.gov/employer/randomizationfaqs.html.

SSA. Social Security Online - History - Frequently Asked Questions, 2011c. http://www.ssa.gov/history/hfaq.html.

SSA. Social Security History: A Myth About Social Security Numbers, 2011d. http://www.socialsecurity.gov/history/ssnmyth.html.

F. Stutzman. An evaluation of identity-sharing behavior in social network communities. In *Proceedings of the 2006 iDMAa and IMS Code Conference, Oxford, Ohio*, 2006.

L. Sweeney. SOS Social Security number watch, 2004. http://privacy.cs.cmu.edu/dataprivacy/projects/ssnwatch/index.html.

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

R. Thomas and J. Martin. The underground economy: Priceless. *;LOGIN:*, 31 (6): 7–16, 2006.

L.H. Thompson. A new role for the social security card. Immigration Control. General Accounting Office. GAO/HRD-88-4, 1988.

United States Department of Justice. Retail hacking ring charged for stealing and distributing credit and debit card numbers from major U.S. retailers: More than 40 million credit and debit card numbers stolen, August 5 2008. http://www.usdoj.gov/opa/pr/2008/August/08-ag-689.html.

U.S. Census Bureau. Table 1. General Mobility, by Region, Sex, and Age: March 2000-2001, 2003. http://www.census.gov/population/socdemo/migration/cps2001/tab01.pdf.

U.S. Census Bureau. Table 4: Cumulative Estimates of the Components of Population Change for the United States, Regions, and States: April 1, 2000 to July 1, 2007, (NST-EST2007-04), 2007. http://www.census.gov/popest/states/tables/NST-EST2007-04.xls.

W. H. Ware. Records, computers and the rights of citizens. Report of the Secretary's Advisory Committee on Automated Personal Data Systems. US Dept of Health, Education, and Welfare and United States, 1973.

R. Wessmiller. Using audit software and the Death Master file to catch crooks. ISACA Newsletter, 2002. http://www.isaca-washdc.org/pages/articles/article-sep2002a-print.htm.

Table 1: Fixed effects model of prediction accuracy (1973-2003, 50 states).

VARIABLES	Accuracy	Accuracy (1-year lagged model)
Births	-1.37e-06***	-1.40e-06***
	(2.36e-07)	(2.29e-07)
Immigrants	1.14e-07	8.69e-08
	(1.11e-07)	(1.07e-07)
EAB	0.175***	0.0968***
	(0.0107)	(0.0131
DividendAct	0.0369***	0.0370***
	(0.0110)	(0.0107)
TaxAct	-0.000244	0.00201
	(0.0154)	(0.0149)
FamilyAct	0.122***	0.118***
	(0.0151)	(0.0147)
EAB-logged		0.108***
		(0.0109)
Constant	0.110***	0.113***
	(0.0187)	(0.0182)
Observations	1550	1550
Number of state_n	50	50
R-squared	0.612	0.636
State FE	YES	
Year FE	YES	

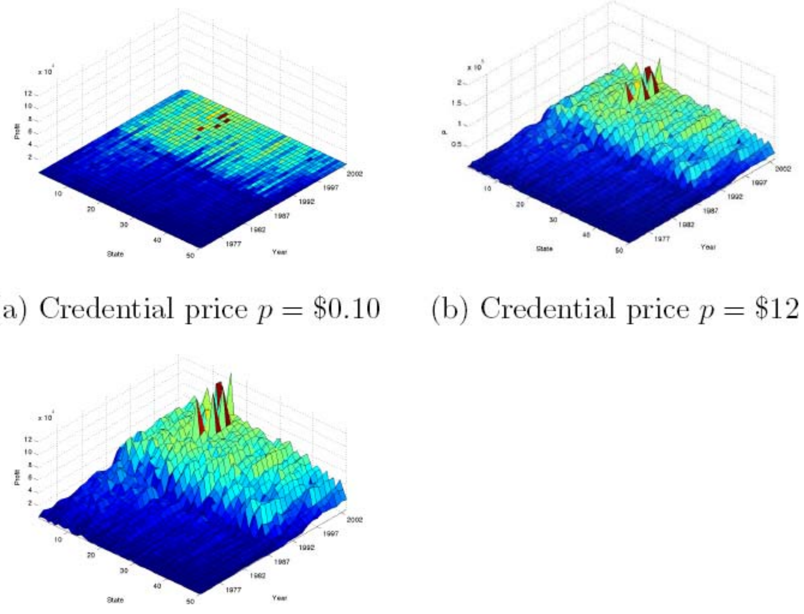
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2: Percentage of SSN records' first five digits accurately predicted at first attempt.

OSN sample	1986-1990	PA, 1986	PA, 1989
DMF Experiment	11%	8%	38%
Number of records	*85,204*	*862*	*566*
OSN Experiment	6%	0%	17%
Number of records	*621*	*37*	*18*
No auxiliary knowledge	0.0014%	0.0014%	0.0014%
Knowledge of state of SSN application	0.012%-1%	0.019%	0.019%

Table 3: Parameter values used in the computations of profits from brute-force SSN predictions attacks.

Variable	Variable Name	Value
Year of birth	*year*	1973-2003
# IP Addresses	*nIp*	10,000
# Independent CC Issuers	*nCra*	20
Max. # Incorrect Attempts Allowed	*k*	3
Birth Data Obtained Percentage	b_p	0.5
CRA Match Percentage	c_p	0.9
Credential Price	*p*	$0.10, $12, $25
Botnet Cost (per day and thousand IPs)	*c*	$100



(a) Credential price $p = \$0.10$ (b) Credential price $p = \$12$

(c) Credential price $p = \$25$

Figure 1: Profits from selling SSNs obtained using a brute force prediction attack at various prices: (a) \$0.10, (b) \$12, and (c) \$25 (see Table 3 for the other parameter values used in this model).

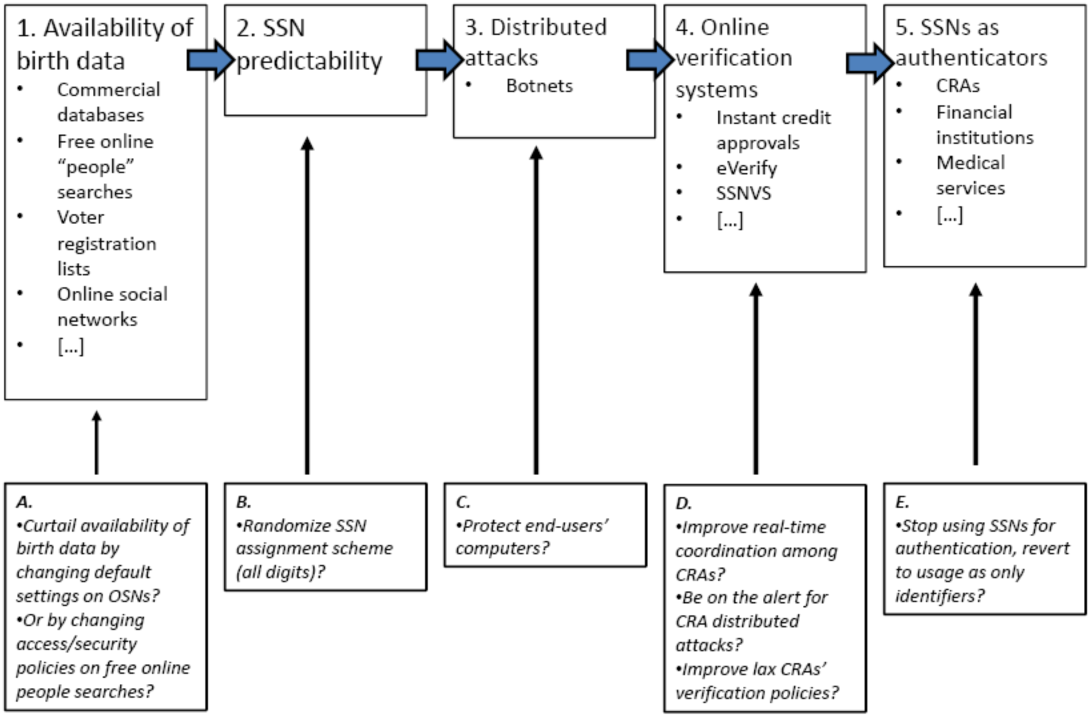
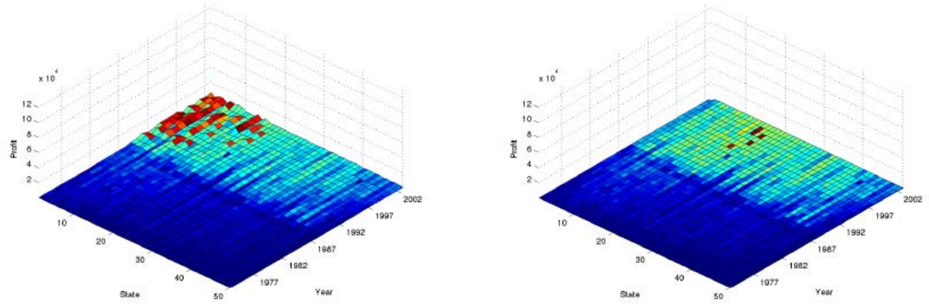

1. Availability of birth data
• Commercial databases
• Free online "people" searches
• Voter registration lists
• Online social networks
• [...]
2. SSN predictability
3. Distributed attacks
• Botnets
4. Online verification systems
• Instant credit approvals
• eVerify
• SSNVS
• [...]
5. SSNs as authenticators
• CRAs
• Financial institutions
• Medical services
• [...]
A.
•Curtail availability of birth data by changing default settings on OSNs?
•Or by changing access/security policies on free online people searches?
B.
•Randomize SSN assignment scheme (all digits)?
C.
•Protect end-users' computers?
D.
•Improve real-time coordination among CRAs?
•Be on the alert for CRA distributed attacks?
•Improve lax CRAs' verification policies?
E.
•Stop using SSNs for authentication, revert to usage as only identifiers?

Figure 2: Vulnerabilities and counter-strategies in the US identity infrastructure.



(a) (b)

Figure 3: (a) Profits from selling SSNs obtained by using the Acquisti-Gross algorithm to predict likely value ranges and verified using botnets and CRAs in the case of real-time communication between CRAs. For this plot we assume a cost per credential of \$25. Profits are markedly reduced from the original model shown in Figure 1. (b) Profits when the value of credential is low (from Figure 1, (a)).

APPENDIX

Social Insecurity: The Unintended Consequences of Identity Fraud Prevention Policies

In this Appendix we present additional details about various technical issues highlighted in the manuscript.

1 Availability of birth data for US residents

In this section we provide additional details about the availability of birth data for US residents from various services. We note that none of the services we tested made SSNs available.

1.1 Data aggregators

Data aggregators offer access to millions of birthday and/or personal records for negligible prices (or, in some cases, for free).[28] Table 1 lists a small subset of the commercial services.

We subscribed to two such services which advertised unlimited searches: http://www.usa-people-search.com and http://www.peoplefinders.com. Since it is difficult to gauge the accuracy and the actual scope of the information contained in the databases associated with the above services, we devised a number of back-of-the-envelope tests. The results should be considered suggestive evidence rather than statistically significant findings.

[28] For instance, usa-people-search states: “Our data is comprised of thousands of public records databases from across the country. We have literally billions of records, all instantly accessible. We compile data from white pages, county records, property records, and MANY other sources” (at http://www.usa-people-search.com/terms.aspx).
peoplefinders.com claims to have records for “almost every adult in the United States in our system.”
usa-people-search.com claims that their “data is comprised of thousands of public records databases from across the country. We have literally billions of records, all instantly accessible.”

First, we manually tested the accuracy of the records provided by the two services against each other, finding very high degree of correspondence, and then also against the database of dates of birth we mined from the online social networking site (OSN sample, see main text), where we also found a high degree of correspondence (for users that we were able to link unambiguously between the data sources). In the majority of cases (17 out of 20, or 85%) the dates of birth were identical. In a small number of cases (3 out of 20) only month and birth year matched, which in many cases is sufficiently accurate for our SSN prediction algorithm. In other cases, no record for the university student could be found in any of the services we used. This happened more frequently for the youngest individuals in our data, suggesting that the services primarily source their information from public records, and therefore the population they capture tends to be older (past college age). This also implies that data from these services complements the population represented in online social networks, which is more frequently of high school and college age.

Second, we tried variations of names of individuals whose personal data we knew to be available from those services, and we verified that no record would be produced under incorrect variations of names. In fact, incorrect versions of names known to be in the database returned no results, indicating that the underlying database has not been populated with synthetic (that is, fake) data. Third, we attempted to estimate the breadth of the data with a very rough back-of-the-envelope test. We considered a particular last name - for instance, “Martin.” According to the 1990 census, there were 678,951 individuals with this last name in the US, representing 0.27% of the population. If we assume (heroically, as an order of magnitude approximation and simplification) that this proportion stays approximately constant across years and sub-populations of the whole US (i.e. states), we can estimate the expected number of people born in a given year with this name using natality statistics. In the year 1980, there were 156,823 births in Pennsylvania. We would therefore expect to find 423 people born in 1980 with the last name “Martin” in PA. Searches on peoplefinders.com for "Martin"s born in 1980 in PA returned 531 matches. Since the search actually returns people who *resided* in PA at some point in their lives (but who were not necessarily born there) one would expect the number of matches estimated based on the birth statistics to be a lower bound for the actual number of matches. We found similar results for all the other names and states we tested - such as "Brown" in Colorado (estimated 308, found 494), or "Thomas" in Iowa (estimated 148, found 209). This suggestive evidence hints at the possibility that the people search engines provide a decent coverage of the population.

The following algorithm shows how an attacker could mine data from the above services:

1. Take the *n* most frequent last names.
2. For each last name and for each state search for people born on a specific date, iterating the date day by day through the desired date range.
3. Download the resulting entries.

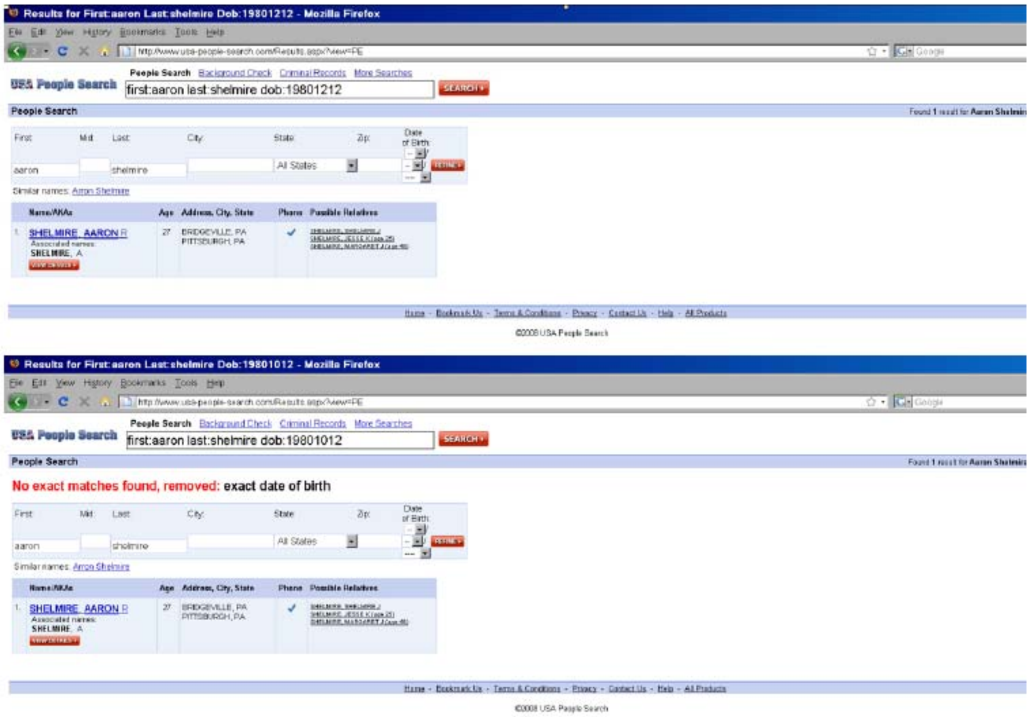
Overall, we found that these services seem to cover a significant portion of the US population and that they provide consistent and non-bogus data. Although both the services we subscribed to explicitly disallow "automated queries" (usa-people-search.com) and "unauthorized use of any robot, spider, or other automated device" (peoplefinders.com), an attacker can easily write mining scripts which imitate human browsing and may go undetected when mining those services for vast amounts of birth data.

Table 1: List of data aggregators with the number of individual birth records they (claim to) have available along with access prices.

Service Name	Number of Records	Cost
unlimited-records.com	80 Million	$59.64/year
peoplefinders.com	Millions	$19.95/month
usa-people-search.com	Millions	$39.95/month
birthdatabase.com	120 Million	Free
netdetective.com	211 Million	$29/3 years
birthrecords.ws	Not disclosed	$22.99/5 years
intelius.com	Not disclosed	$19.95/24h

1.2 Online people searches

As noted in the main article, certain online people services allow for the advanced search of individuals by name and dates of birth. Since sites such as http://www.usa-people-search.com or http://www.zabasearch.com do not use CAPTCHA systems (see http://www.captcha.net/) to prevent their abuse by automated scripts, repeated variations of names and dates of birth can be issued to ping and test an individual’s actual name, as in the example presented in Figure 1. Here we show screen captures of a search using the exact date of birth of an individual (first row) and a search with an incorrect date of birth of the same individual (second row). It is trivial to write automated scripts which run brute force variations of an individual’s date birth, and by comparing the answers that the system provides to the queries, eventually pinpoint his or her exact date of birth.



Results for First:aaron Last:shelmire Dob:19801212 - Mozilla Firefox

http://www.usa-people-search.com/Results.aspx?View=PE

USA People Search

first:aaron last:shelmire dob:19801212

People Search

Found 1 result for Aaron Shelmir

First	Mid	Last	City	State	Zip	Date of Birth
aaron		shelmire		All States		

Similar names: Arron Shelmire

	Name/AKAs	Age	Address, City, State	Phone	Possible Relatives
1.	SHELMIRE, AARON R Associated names: SHELMIRE, A	27	BRIDGEVILLE, PA PITTSBURGH, PA	✓	SHELMIRE, SHELMIRE J SHELMIRE, JESSE K (age 25) SHELMIRE, MARGARET J (age 48)

Home - Bookmark Us - Terms & Conditions - Privacy - Contact Us - Help - All Products

©2008 USA People Search

Results for First:aaron Last:shelmire Dob:19801012 - Mozilla Firefox

http://www.usa-people-search.com/Results.aspx?View=PE

USA People Search

first:aaron last:shelmire dob:19801012

People Search

Found 1 result for Aaron Shelmir

No exact matches found, removed: exact date of birth

First	Mid	Last	City	State	Zip	Date of Birth
aaron		shelmire		All States		

Similar names: Arron Shelmire

	Name/AKAs	Age	Address, City, State	Phone	Possible Relatives
1.	SHELMIRE, AARON R Associated names: SHELMIRE, A	27	BRIDGEVILLE, PA PITTSBURGH, PA	✓	SHELMIRE, SHELMIRE J SHELMIRE, JESSE K (age 25) SHELMIRE, MARGARET J (age 48)

Home - Bookmark Us - Terms & Conditions - Privacy - Contact Us - Help - All Products

©2008 USA People Search

File Edit View History Bookmarks Tools Help
People Search Background Check Criminal Records More Searches
SEARCH
REFINE
VIEW DETAILS
File Edit View History Bookmarks Tools Help
People Search Background Check Criminal Records More Searches
SEARCH
REFINE
VIEW DETAILS

Figure 1: Screen captures of searches with an online people search engine using the correct (first row) and incorrect DOB (second row), enabling automatic determination of the DOB.

Some of the sources listed in the previous sub-sections do *not* provide information about the individual’s state of birth but rather his history of residences, which often includes the individual’s state of birth. While US citizens tend to be very mobile (especially in the 20 to 35 age bracket), most of the mobility which takes place every year occurs within the same state where a person was already residing. For instance, according to the US Census Bureau, between 2000 and 2001 only 3.4% of the US population moved to another State (most moves are within the same county and within the same State) (U.S. Census Bureau 2001). Although such moves accumulate over the years, the probability is high that the address history reported by these services (and, in some services, the addresses of relatives) will include the state of birth of the individual.

2 Verification channels

In this section we extend the discussion of the feasibility and limitations of using various online channels to verify SSN predictions. We emphasize that the verification strategies described below are informed inferences based on publicly available data about those channels and the attack strategies that criminals are known to be using.

2.1 Verification channels: Computational Cyber Attacks

Recent years have witnessed the emergence of cybercriminal networks and the raise of "computational" attacks. Cybercriminal networks are now based on the division and specialization of labor (Anderson et al. 2008, Moore et al. 2009) and the exploitation of "botnets" (Cooke et al. 2005) of compromised hosts to engage in brute force attacks. Such attacks have become quite common in various forms thanks to the commoditization of tools of attack (Anderson et al. 2008, Moore et al. 2009) and low botnets rental prices.[29] Botnets have been used in computational attacks for spam and phishing (see (Jakobsson and Myers 2006)); to harvest online résumés for personal information (Anderson 2007); to break CAPTCHAs verification systems (see http://www.captcha.net/) and create mass amounts of new email accounts (Websense 2008); to spam online social networks (Leyden 2008b); to conduct phishing attacks on social networking sites (Symantec 2008); and, in fact, to create mass amounts of fake profiles on social networking sites (Leyden 2008a) with the goal of spamming or harvesting personal information.

[29] Industry data cited in (Moore et al. 2009, p.3) suggests that 10 million computers are infected with malware designed to steal online credentials. An exact estimate of the number of bots currently available to cybercriminals is hard, as old botnets get taken down while new ones appear (Cho et al. 2010). In recent years, botnets such as BredoLab, Mariposa, and Conficker have been claimed in the security press to control from 10 to 30 million bots (see, e.g., Mullins (2010)).

In the case of Social Security numbers, attacker can exploit online verification channels as "oracle machines" (Papadimitriou 1994) to find which SSN corresponds to an individual with a given birth date. The attacker repeatedly queries the oracle in a brute force approach similar to a cryptographic *dictionary* attack.[30] Identity theft criminal rings have already engaged in a practice called "tumbling" that consists of starting with an SSN known to be correct, then "changing one digit at a time for each new persona" (Weston 2008)) in the attempt to create more fake identities (ID Analytics 2005, p. 12). At least two alternative strategies of exploiting algorithms that predict Social Security numbers are possible. One strategy consists of cycling through as many variations of each target's SSN as the verification channels allow the attacker to test, while distributing the variations across channels and, at each channel, across accounts the attacker has been able to compromise. An alternative strategy consists of trying out very few variations per target victim, but move quickly through vast amounts of targets. This would lower the number of target SSNs the attacker will be able to identify, but also the probability of being detected.[31]

2.2 Verification channels: The SSA's SSN Verification Service (SSNVS)

Under the Privacy Act of 1974 (P.L. 93-579), the SSA can disclose an individual's SSN to third parties with that individual's consent and, in a number of controlled scenarios, also without his consent. In fact, as noted in the paper, the SSA offers an online service that allows companies and self-employed individuals to verify other individuals' SSNs based on their names and DOBs (SSA 2007).[32] The service is actually offered to *avoid* SSN fraud: representatives of legitimate employers armed with knowledge of their company's Employer Identification Number (EIN), or

[30] An example of a brute force approach used in Acquisti and Gross (2009) has the attacker try out a target's predicted SSN before moving up and down the Serial Numbers in 1-integer steps for the following attempts, while keeping the predicted first five digits constant. This approach guarantees that roughly 1 out of 10 SSNs issued between 1988 and 2003 will be found with fewer than 1,000 attempts.

[31] As discussed in (Herley and Florêncio 2008a), "Far more likely to succeed is a bulk guessing attack against a large number of accounts [...] Since only a small number of unsuccessful logins are attempted at any individual userID the attack is far harder to detect."

even self-employed individuals with legitimate SSNs, can verify the SSNs of prospective employees by phone, by letter, and online - receiving responses about the validity of up to 250,000 SSNs in one night (SSA 2007). Access to the service is free, but requires previous registration and vetting by the SSA.[33] However, the vetting process may be compromised: an attacker may use stolen but verified identities (using companies' EINs or simply self-employed individuals' SSNs) or even set up front companies[34] to attempt to register for the service. The attacker would then pose as an organization employee or a self-employed individual hiring new employees, and submit requests for verification of predicted SSNs - basically harvesting stolen SSNs to produce more stolen SSNs. Indeed, (Westat 2007, p.xxvi) reports that the Social Security Administration has experienced attempts by third parties to pose as employers and get access to SSNVS data.

Through the SSNVS service, the SSA sends back messages confirming the identification of a correct SSN or noting an error in the process (SSA 2007). Importantly, the SSA also keeps track of incorrect attempts at verifying a certain identity. A cunning attacker, however, would distribute verification requests of permutations of a set of predicted SSNs over time and over several stolen (and previously verified) identities. Within each verification batch, the attacker would properly mix SSNs known to be valid with those he actually wants to verify - in order to lower the ratio of incorrect SSNs in any such attempt, and make the attack less detectable.

A similar but more limited service is also provided by the US Selective Service System through the Selective Service Online Registration Verification.[35]

[32] The service is available at http://www.ssa.gov/bso/bsowelcome.htm.

[33] See http://www.socialsecurity.gov/employer/ssnvs_handbk.htm#regin.

[34] ID thieves can easily set up front organizations to obtain the information they need. ChoicePoint was conned by a criminal organization which opened several accounts with the data broker in 2005 to purchase individual reports "typically cost[ing] between $5 and $17" (O'Harrow Jr 2005).

[35] Available at https://www.sss.gov/RegVer/wfVerification.aspx. Almost "all male US citizens, and male aliens living in the US, who are 18 through 25, are required to register with Selective Service" as discussed in (Selective Service System 2008).

2.3 Verification channels: The DHS's E-Verify Service

Similarly to the SSNVS, the Department of Homeland Security's E-Verify system is an Internet-based program that allows the verification of SSNs. It is operated by the US Citizenship and Immigration Services in partnership with the SSA.[36] It is free to any employer who registers for the service. It provides "an automated link to Federal databases to help employers determine employment eligibility of new hires and the validity of their Social Security numbers, [...] virtually eliminating Social Security mismatch letters" (Department of Homeland Security 2007, p. 1). Every employer who registers as a user receives a user ID and a password.[37] An attacker who succeeded in impersonating an employer, or who could steal a legitimate employer's password, may therefore abuse the system for SSN verification purposes. While SSNVS has developed some protection against such attacks, anybody wanting to access E-Verify can pose as an employer and gain access simply by signing a Memorandum of Understanding (Westat 2007, p.xxvi), (GAO 2007). According to a GAO report, "taking actions to ensure that employers are legitimate when they register for E-Verify is a long term goal for the program" (GAO 2007). However, the GAO also notes that implementing such controls require access to information from other agencies to which USCIS currently does not have access. Responses to queries to the system (which can consist simply of an individual's name, date of birth, and presumptive SSN) are sent within seconds of submitting the query on the Internet.[38] Responses may consists of "employment authorized" (which verifies the SSN to the attacker, unless data errors in the system originate false positive results: see (Electronic Privacy Information Center 2007)); "DHS verification in process," implying that the SSN matched SSA records, but the work eligibility information did not match

[36] The program is available at https://www.vis-dhs.com/EmployerRegistration/StartPage.aspx? JS=YES.

[37] According to (Electronic Privacy Information Center 2007), at least 200,000 Federal contractors will be required to register for the system. However, as noted above, any employer - not just those who are Federal contractors - may sign up for the service.

[38] There exists also a batch access method to verify multiple records at the same time, see (Department of Homeland Security 2007).

DHS's records; and "SSA tentative nonconfirmation," implying that the SSN could not be verified (Department of Homeland Security 2008).

One of the challenges is the sheer size of the program: if participation were made mandatory for all employers, "the program would have to accommodate all of the estimated 7.4 million employers in the United States" (GAO 2007, p. 10), raising the issue of how to efficiently authenticate millions of employers (including small businesses) without making it possible for criminals to impersonate small firms, access the program, and try to use it to test SSNs by distributing verifications across different accounts. Furthermore, it has been projected that 63 million queries per year could be submitted to E-Verify if the program were made mandatory (GAO 2007, p. 10), making it particularly difficult to detect "low noise" attacks based on a few thousand attempts. Furthermore, As of March 2011, the Department of Homeland Security launched a "self-check" program to U.S. residents to use E-Verify in order to test their own eligibility for employment before looking for jobs.[39]

2.4 Verification channels: Instant Credit Approval Services

In the private sector, numerous online "instant credit card application" services (or similar wireless carriers and "instant" lending companies verification services (ID Analytics 2005)) take as input an individual's name, date of birth, and SSN (plus, sometimes, auxiliary information such as his address or phone number) and return as output a preliminary approval or denial of that individual's application for credit or service.[40] This response provides the attacker knowledge about the accuracy of his SSN predictions. These services belong to the "faceless, instant" family of applications that the Internet has made popular, and are particularly vulnerable to attacks, because

[39] See http://edition.cnn.com/2011/US/03/21/worker.eligibility/.

[40] Auxiliary information is not necessarily validated during the verification process, even when it may be requested as part of it. For instance, the passage of the Drivers Privacy Protection Act (P.L. 103-322), originally enacted in 1994, has made it harder for

in instant credit application lenders often do not have as much time to properly verify the accounts (ID Analytics 2003, p. 13).[41]

2.5 Verification channels: Phishing

In a phishing attack, the criminal attempts to lure victims into revealing sensitive personal information, for instance by sending them fake messages from banks or organizations the victims may have an account with. "Spear" phishing attacks (Jakobsson and Myers 2006), based on social engineering and targeted at the individual with personal and contextual clues that give legitimacy to the phishing message, have been proved to be particularly successful at baiting unsuspecting victims into providing personal data, such as accounts' passwords (Jagatic et al. 2007). Since we showed individual ANGNs to be highly predictable for specific cohorts, a phishing email that contained an individual's first five SSN digits and asked him to log onto a (fake) site (such as a site spoofing http://www.irs.gov) by verifying the remaining digits, in order to avoid a penalty or receive a refund, may appear legitimate and lure individuals into providing the remaining digits.[42]

2.6 Verification channels: Additional attack vectors

Additional vectors of attack also rely on exploiting knowledge of the first digits of an SSN and auxiliary personal information of the individual in order to gain access to the complete SSN through social engineering; or rely on more traditional hacking strategies to find and harvest the last four digits of several individuals' SSNs:

organizations to purchase Driver License number information - and therefore to validate Driver License numbers in real time in faceless online applications.

[41] According to (ID Analytics 2003, p. 11), instant credit applications result in higher fraud rates than non-instant credit applications, as well as in more rapid spending of the credit fraudulently obtained. Furthermore, certain services explicitly advertise the possibility of getting credit cards for people regardless of [their] credit history. See, for instance, http://www.creditcardguide.com/index_needcredit.html.

[42] Unrelated examples of scams based on IRS and SSN data have already been reported (Internal Revenue Service 2008).

- Spoofing electronic connections to intercept the digits of an SSN that are used for authentication in numerous services (such as cell phone providers, voicemails, or library logins).

- Calling banks or financial services and using the information already available to the attacker - such as phone number, address, and some of the digits of the SSN - in order to learn the remaining digits - also known as "pretexting" (The President's Identity Theft Task Force 2007).

- Accessing many public records that still include an individual's last four digits of the SSN as identifiers - for instance in the military[43] and other sectors.[44] As noted in the article, various states' recent legislative initiatives have been directed at protecting SSNs.[45] Unfortunately, most have done so by restricting the public usage of *only* their first five digits, and allowing the last four to be associated to individual names in public documents. For instance, a law recently passed in Nevada states that "the last four digits of a Social Security number are not personal information for the purposes of these provisions" (A.B. 600, Signed by Governor 6/4/07, Chapter 324). As discussed in the main body, in light of how predictable we found the first five digits of someone's SSN to be, such legislations may be misguided.

- Hacking and accessing non-public records held by several companies and organizations - from wireless carriers to financial institutions - that store the last four digits of individuals' SSNs for authentication purposes (see e.g. (Irwin Financial Corporation 2007)).

[43] See http://www.stripes.com/article.asp? section=104&article=62242&archive=true.

[44] For instance, for NY attorneys, the Local Civil Rule 11.1(b) (Eastern District and Southern District NY) formerly claimed: "Every pleading, written motion, and other paper that is signed by an attorney must show directly after the typed name of the attorney (1) the initials of the attorney's first and last name, and (2) the last four digits of the attorney's social security number, or any other four-digit number registered by the attorney with the clerk of the court." See also http://www.njd.uscourts.gov/cm-ecf/ECFREG.pdf, where the NJ attorney ID is referred to as the combination of the initials of the first and last name and the last four digit of his SSN.

[45] See http://www.ncsl.org/programs/lis/privacy/SSN2007.htm.

2.7 Verification strategies

As noted in the main article, at least two alternative strategies of exploiting the verification channels we described above seem plausible.

One strategy consists of cycling through as many variations of each target's SSN as the verification channels allow the attacker to test, while distributing the variations across channels and, at each channel, across accounts the attacker has been able to compromise.[46] An attacker may limit himself to testing out, for instance, 1,000 variations of a victim's SSN across several different channels, before abandoning that target altogether. The attacker would not know in advance *which* victims will be identified within 1,000 attempts, but can calculate the odds of identifying a certain number of victims' SSNs. Naturally, the attacker would not have to try out 1,000 SSNs for each potential victim: for many of them he will hit the right SSN well before 1,000 attempts.

An alternative strategy consists of trying out very few variations per target victim, but move quickly through vast amounts of targets. This would lower the number of target SSNs the attacker will be able to identify, but also the probability of being detected. Such alternative strategy can be explained with the observation that the statistics presented in this paper (for instance, "33.66% of SSNs issued in New Mexico in 1996 may be matched with fewer than 1,000 attempts") may be interpreted in two different ways: first, if the attacker tried, for instance, 1,000 attempts on each SSN issued in New Mexico in 1996, he may expect to identify 33.66% of them; second, if the attacker tried to predict 1,000 random different SSNs issued in New Mexico in 1996, on average he would identify 0.03% of them at the first attempt.[47]

[46] As noted above, in a brute force approach the attacker would try out a target's predicted SSN before moving up and down the SNs in 1-integer steps for the following attempts, while keeping the predicted ANGN constant.

[47] In reality, the actual number is not necessarily one thousandth of the accuracy rate over 1,000 attempts, since the accuracy rate does not necessarily progress linearly. In supplementary tables linked from the On-line Supplement we provide prediction accuracy ratios for numbers of attempts ranging from as low as 1 or 10 to as high as 1,000 or 5,000. For instance, nationwide, the weighted mean of the percentage of whole SSNs for DMF records with post-1988 dates of birth which can be matched with fewer than *one hundred* attempts is 1%, with peaks as high as 17% for smaller states. We actually *did* match 1 DMF record's SSN at the first attempt in New

A rational attacker will decide how many variations of the same target's SSN to try out, and how to spread them over time and over verification channels and accounts, as function of a number of factors: the availability of botnets to hide and distribute simultaneous attacks, the threshold which triggers the verification system's alert, and the sophistication of the alert and defense mechanisms employed by the verification channel. Together, these parameters will determine how long and how heavily the attacker may be able to exploit (and, in fact, combine together) the first and the second strategies.

Within each verification batch, the attacker would properly mix SSNs known to be valid with those he actually wants to verify - in order to lower the ratio of incorrect SSNs in any such attempt, and make the attack less detectable. For instance, in order to engage in 1,000 verification attempts on each victim's SSN for a large number of victims, the attacker would need to exploit botnets and spread his efforts through different channels, combining several online instant approval services and a few hundred compromised accounts at services such as SSNVS and E-Verify, and testing only a handful of variations of every target's SSN at each service. Assuming that five attempts were sufficient to halt the verification of an identity at a certain service provider, and assuming that the victim's SSN can be matched within 1,000 attempts, an attacker would have to distribute the attack across 200 services - or, viceversa, distribute the attacks over time since alarms on a certain account cannot permanently freeze his account: unlike passwords, SSNs cannot be revoked after unsuccessful attempts to log into a system. This creates a trade-off for the attacker between choosing many attempts spread over a short period of time, or more evenly spread attempts. In reality, since SSNs are not revokable, it is not clear that a service provider (or the SSA itself) could actually block any account; more likely, if a botnet built by an attacker kept testing variations of an SSN, the service being queried could blacklist its IP address - and the attacker would then move to

Mexico in 1996. In general, we found that we could match with fewer than 10 attempts 0.01% of New Mexico records with 1996 the next bot under his control in the botnet in order to keep verifying variations of SSNs for that identity.

2.8 Defenses and Potential Vulnerabilities in Verification Channels

All of these verification channels implement defenses against attempts of abuse and brute force attacks. Some of those defenses are well known (such as blocking IP addresses which have originated suspicious requests or failed logins), and others are jealously guarded, in order not to provide attackers the advantage of actionable intelligence. However, it is possible to gain some understanding of the vulnerability of those mechanisms from publicly available documents, as well as from empirical observation of the activities of credit fraud criminal rings and their increasing sophistication in using knowledge about the screening mechanisms of the financial industry to bypass its defenses,[48] exploiting its highly distributed and relatively unregulated nature. We will focus on the abuse of online credit verification systems.

First, different credit services which cater directly to the end consumer are likely to query the same credit reporting agencies' databases (such as those by Experian, Equifax and TransUnion) to verify data provided by applicants on their websites. However, services offered by the different agencies and various financial institutions do not communicate in real-time.[49] This implies that awareness of repeated invalid credit requests for the same identity does not propagate quickly

dates of birth (out of 101 DMF records in that year and state).

[48] See (McCarty 2003, ID Analytics 2005, Thomas and Martin 2006). Indeed, "credit" hackers have been described by (Spencer 2008). For instance, "[i]n at least one case, identity thieves made the minimum payments on the accounts so that the credit cards would continue to be active" (Hoofnagle 2007, p.110). On the other hand, also see (Herley and Florêncio 2008b) on the possibility that phishing attacks may not prove particularly efficient for most low-skill hackers.

[49] IDAnalytics discovered that fraud rings perform cluster attacks with clear temporal patterns (ID Analytics 2005, p. 3), taking advantage of the lack of timely exchange of application data within the industry. IDAnalytics therefore denounced that the lack of cooperation within the industry makes it harder to reduce identity fraud at the point of application (ID Analytics 2003, p. 13): for instance, "[h]its and flags [on potentially fraudulent accounts] may take 90 days or longer to appear in known fraud databases, while many cases of identity fraud occur in a concentrated period of 30 to 60 days." (ID Analytics 2003, p. 9).

across different market players,[50] allowing an attacker to distribute his verification attempts across different channels.

Second, in such an environment, thoughtful attackers can choose whether to cluster several attempts to identify a set of targets' SSNs in a short period of time, but spreading them across different channels and accounts at those channels; or to spread them both over time and across channels, in order not to alert built-in security mechanisms at each specific institution (which may halt the application for a certain combination of individual name and date of birth after a given number of failed tries). In the case of online instant credit approval, for instance, numerous services are available: during an independent analysis of such services on the Internet that we ran for this study, we counted more than 70 of them (plus many more sites offering duplicate services under different names).[51]

Third, and perhaps most importantly, abundant empirical evidence points at the laxness of credit bureaus in properly screening credit applications (Hendricks 2003, ID Analytics 2005, 2006, Hoofnagle 2007). Because consumer credit reports are known to contain, inevitably, errors and inconsistencies, credit reporting agencies do accept as valid applications with incorrect information - including names, addresses, and even SSN digits (FTC 2004). For instance, CRAs "have mismerged data about two different consumers because their algorithms tolerate what's known as 'partial matches''' (Hendricks 2003, p. 3),[52] and "[i]n certain circumstances, some CRA algorithms tolerate a partial SSN match of 7 out of 9 digits" (Hendricks 2003, p. 4). In fact, while in "current account" frauds the attacker, to gain access to an account *already* created and owned by the

[50] For instance, a mortgage loan fraudulent activity reported by financial institutions consists of "[b]orrowers sign[ing] multiple mortgages on the same property from multiple lenders. The mortgage settlements were held within a short period of time to prevent the lenders from discovering the fraud" (Financial Crimes Enforcement Network 2006).

[51] See, for instance, http://www.creditcards.com/instant-approval.php.

[52] For instance, the SSN of Judy Thomas, a resident of Klamath Falls, Oregon, differed from the SSN of Judith Upton, of Stevens, Washington, only by one digit. "This, probably coupled with partial matches on first name, caused CRA's algorithm to assume that the one-digit difference was a clerical error and that Thomas and Upton were the same person, with one SSN. Many of Upton's derogatory trade lines were improperly merged on to Thomas' credit report, causing delays in obtaining a mortgage and other hassles and distress" (Hendricks 2003, p. 3).

individual, needs not just the victim's name, date of birth, and SSN, but (most often) also additional passwords or personal information, in "new account" frauds, the attacker more likely only needs to use the victim's name, date of birth, and SSN to create a *new* account on the victim's name. Therefore, new account frauds can be perpetrated even without knowledge of the victim's address, phone number, or other pieces of personal information. Mounting empirical evidence suggests, in fact, that providing an SSN and a date of birth which match that SSN is sufficient to create *new* fraudulent accounts (Cook 2005, Hoofnagle 2007, Samuelson Law, Technology & Public Policy Clinic 2007), even when the name associated with that SSN did *not* match,[53] or the address was wrong,[54] or even - as noted above - some of the submitted SSN digits were wrong.[55] In other words, financial institutions often authenticate SSNs only by comparing them to the associated dates of birth, instead of actually vetting that the numbers are issued to the correct individuals (Hoofnagle 2007, Cook 2005). As noted by (Samuelson Law, Technology & Public Policy Clinic 2007, p.2), several lawsuits against credit issuers for their negligence in opening new accounts to criminals document a lack of reliable controls in the authentication process: businesses which grant credit "do little to ensure names and Social Security numbers match and credit bureaus allow perpetrators to establish credit files using other people's Social Security numbers" (Mitchell 2004, E1).[56]

[53] The rising threat of synthetic identity theft, discussed below, provides additional empirical evidence that credit accounts can be created simply with knowledge of a correct SSN and its associated date of birth, even when the associated name may be incorrect. In synthetic identity theft, SSNs which do *not* match the correct individual's name (but are associated with the correct date of birth) are sufficient to create new credit accounts: see (Hoofnagle 2007, p. 118) and further below.

[54] Instant credit approval applications are routinely accepted even with false addresses and phone numbers - see (ID Analytics 2003, p. 46). See also (Acohido and Swartz 2008, p. 33): "[f]irst, the fraudster obtains Tin Ahern's SSN. Next, the crook applies online for a pay-as-you-go cell phone account under the name Tina A. Hern, using Tina Ahern's SSN but a different billing address. A few months later - after paying the cell phone bill and thus establishing a payment history and new billing address for Tina A. Hern - the thief successfully applies for a credit card using the name Christina Hern, instead of Tina A. Hern or Tina Hern, with two digits of the SSN inverted."

[55] The practice of "tumbling" - which consists of slightly changing numerical details in fraudulent applications, such as addresses and, in fact, the manipulation of known SSNs across multiple account applications - by identity theft criminal rings has been documented by IDAnalytics: see (ID Analytics 2006, p. 24) and (ID Analytics 2005, pp. 16-17).

[56] See also Howard (2005), on a case involving a victim of "negligent enablement of imposter fraud," which happens when a financial institution negligently opens a bank account or extends credit to an identity thief. See also Apodaca v. Discover Financial Services, et al., 417 F.Supp.2d 1220 (D.N.M. 2006), about punitive damages available against credit reporting agency whose systems fail to take into account information provided by consumer in mixed file disputes.

Such laxness may act as a double-edged sword for an attacker: it suggests that a service approval of a given SSN may not guarantee that all the SSN digits have been verified as correct; on the other hand, it may not matter, since such lack of dependable vetting implies that identity theft attempts with predicted SSNs where only seven or eight digits out of nine matched the actual SSN, and auxiliary information were incorrect, could nevertheless be successful.[57] As noted in Acquisti and Gross (2009) the implication is that the SSN prediction accuracies found in the DMF and OSN experiments may be conservative by orders of magnitude: " with just 10 or fewer attempts per target, the inquiries associated with 9.2% of all SSNs issued after 1988 could be accepted as valid by CRAs, and almost 30% of those issued in the 25 states with fewer births."

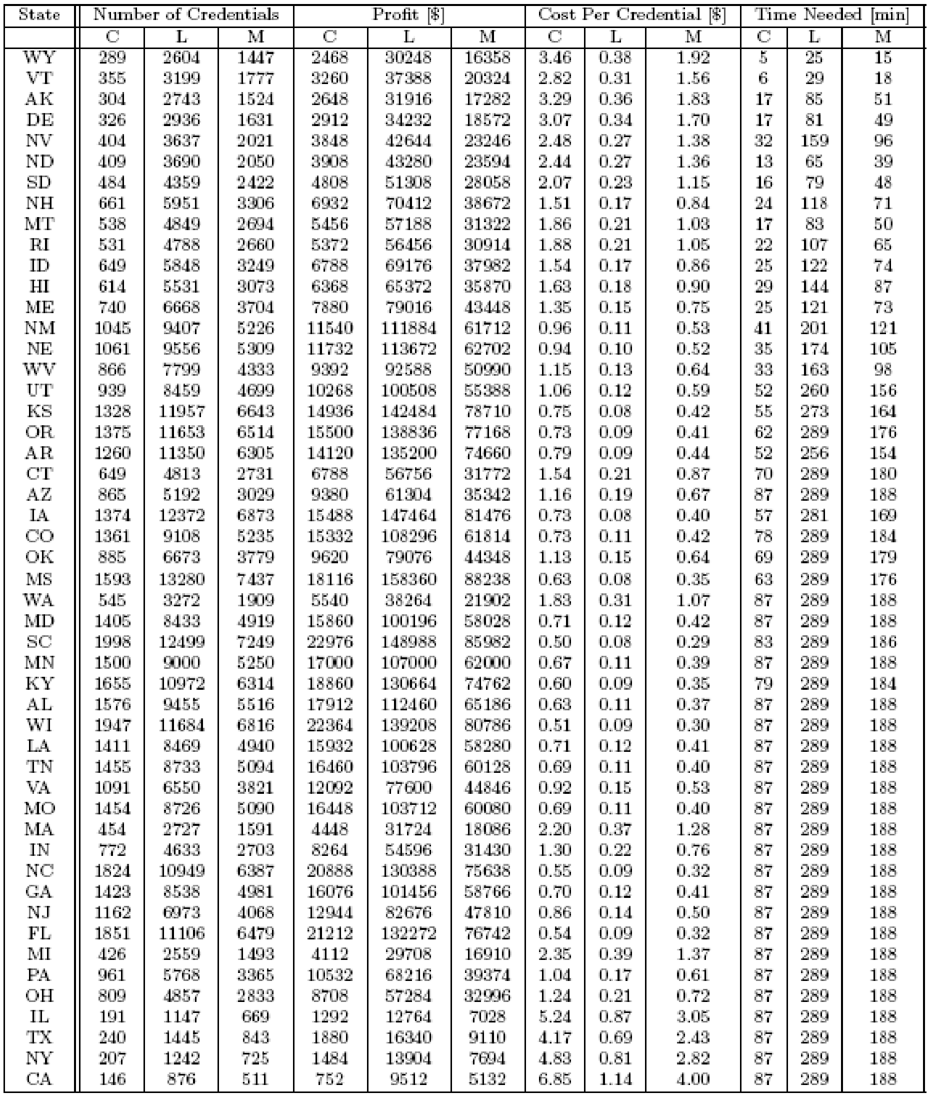
Fourth, and ultimately, the critical vulnerability in the US identity verification infrastructure is due to the peculiar nature of Social Security numbers as passwords: once assigned, SSNs *cannot* be changed to avoid *future* frauds (obtaining a new SSN is difficult even *after* identity fraud (SSA 2008)). Hence SSNs cannot be halted, revoked, or replaced in the same way an ordinary password or credit card (and access to its associated account) can be.[58] This creates an additional layer of difficulty for the protection of digital identities: the physical sources of the attacks (such as the IP addresses originating malicious verification attempts) can be blacklisted, but organized attackers can easily replace them; the *object* of their attack, however, cannot be so easily blocked.

[57] According to (Samuelson Law, Technology & Public Policy Clinic 2007, p.1), "[m]ounting evidence suggests that some credit grantors engage in the practice warned against in the Commission's identity theft materials, albeit with the full SSN: they use the SSN as a password in verifying an individual's identity." The Federal Trade Commission also admitted that "[o]rganizations may use the SSN (*alone* or in combination with other information) to verify identity" (FTC 2007, p.1) (emphasis added by the author). The Commission recently noted that "[a]lthough there is disagreement as to whether a thief can use the victim's name and SSN alone to steal her identity, it is generally understood that, at the least, the SSN facilitates identity theft, i.e., that it is a necessary, if not necessarily sufficient, data element for many forms of this crime to occur" (FTC 2008, p.3). The Commission further acknowledged that experts have highlighted "instances in which credit was granted based on applications full of inconsistencies and mismatched information." (p. 12). On p. 4, (FTC 2008) further noted that "[i]n other cases, businesses may not be requiring the right type of authentication (such as requiring only a name and SSN, or other readily available information, for account access), or their employees may not be following the company's procedures."

[58] Although ineffective against verification attacks conducted through the SSNVS or E-Verify, credit freeze laws can make new account frauds harder to perpetrate. Through these laws, consumers can "freeze" their credit reports and therefore limit potential creditors' (and criminals') access to them (Hoofnagle 2005).

3 Estimates of number of credentials obtained per target and other variables under various parameter assumptions

Estimates of number of credentials obtained per target and other variables under various parameter assumptions are available in Table 2.



State	Number of Credentials			Profit [$]			Cost Per Credential [$]			Time Needed [min]		
	C	L	M	C	L	M	C	L	M	C	L	M
WY	289	2604	1447	2468	30248	16358	3.46	0.38	1.92	5	25	15
VT	355	3199	1777	3260	37388	20324	2.82	0.31	1.56	6	29	18
AK	304	2743	1524	2648	31916	17282	3.29	0.36	1.83	17	85	51
DE	326	2936	1631	2912	34232	18572	3.07	0.34	1.70	17	81	49
NV	404	3637	2021	3848	42644	23246	2.48	0.27	1.38	32	159	96
ND	409	3690	2050	3908	43280	23594	2.44	0.27	1.36	13	65	39
SD	484	4359	2422	4808	51308	28058	2.07	0.23	1.15	16	79	48
NH	661	5951	3306	6932	70412	38672	1.51	0.17	0.84	24	118	71
MT	538	4849	2694	5456	57188	31322	1.86	0.21	1.03	17	83	50
RI	531	4788	2660	5372	56456	30914	1.88	0.21	1.05	22	107	65
ID	649	5848	3249	6788	69176	37982	1.54	0.17	0.86	25	122	74
HI	614	5531	3073	6368	65372	35870	1.63	0.18	0.90	29	144	87
ME	740	6668	3704	7880	79016	43448	1.35	0.15	0.75	25	121	73
NM	1045	9407	5226	11540	111884	61712	0.96	0.11	0.53	41	201	121
NE	1061	9556	5309	11732	113672	62702	0.94	0.10	0.52	35	174	105
WV	866	7799	4333	9392	92588	50990	1.15	0.13	0.64	33	163	98
UT	939	8459	4699	10268	100508	55388	1.06	0.12	0.59	52	260	156
KS	1328	11957	6643	14936	142484	78710	0.75	0.08	0.42	55	273	164
OR	1375	11653	6514	15500	138836	77168	0.73	0.09	0.41	62	289	176
AR	1260	11350	6305	14120	135200	74660	0.79	0.09	0.44	52	256	154
CT	649	4813	2731	6788	56756	31772	1.54	0.21	0.87	70	289	180
AZ	865	5192	3029	9380	61304	35342	1.16	0.19	0.67	87	289	188
IA	1374	12372	6873	15488	147464	81476	0.73	0.08	0.40	57	281	169
CO	1361	9108	5235	15332	108296	61814	0.73	0.11	0.42	78	289	184
OK	885	6673	3779	9620	79076	44348	1.13	0.15	0.64	69	289	179
MS	1593	13280	7437	18116	158360	88238	0.63	0.08	0.35	63	289	176
WA	545	3272	1909	5540	38264	21902	1.83	0.31	1.07	87	289	188
MD	1405	8433	4919	15860	100196	58028	0.71	0.12	0.42	87	289	188
SC	1998	12499	7249	22976	148988	85982	0.50	0.08	0.29	83	289	186
MN	1500	9000	5250	17000	107000	62000	0.67	0.11	0.39	87	289	188
KY	1655	10972	6314	18860	130664	74762	0.60	0.09	0.35	79	289	184
AL	1576	9455	5516	17912	112460	65186	0.63	0.11	0.37	87	289	188
WI	1947	11684	6816	22364	139208	80786	0.51	0.09	0.30	87	289	188
LA	1411	8469	4940	15932	100628	58280	0.71	0.12	0.41	87	289	188
TN	1455	8733	5094	16460	103796	60128	0.69	0.11	0.40	87	289	188
VA	1091	6550	3821	12092	77600	44846	0.92	0.15	0.53	87	289	188
MO	1454	8726	5090	16448	103712	60080	0.69	0.11	0.40	87	289	188
MA	454	2727	1591	4448	31724	18086	2.20	0.37	1.28	87	289	188
IN	772	4633	2703	8264	54596	31430	1.30	0.22	0.76	87	289	188
NC	1824	10949	6387	20888	130388	75638	0.55	0.09	0.32	87	289	188
GA	1423	8538	4981	16076	101456	58766	0.70	0.12	0.41	87	289	188
NJ	1162	6973	4068	12944	82676	47810	0.86	0.14	0.50	87	289	188
FL	1851	11106	6479	21212	132272	76742	0.54	0.09	0.32	87	289	188
MI	426	2559	1493	4112	29708	16910	2.35	0.39	1.37	87	289	188
PA	961	5768	3365	10532	68216	39374	1.04	0.17	0.61	87	289	188
OH	809	4857	2833	8708	57284	32996	1.24	0.21	0.72	87	289	188
IL	191	1147	669	1292	12764	7028	5.24	0.87	3.05	87	289	188
TX	240	1445	843	1880	16340	9110	4.17	0.69	2.43	87	289	188
NY	207	1242	725	1484	13904	7694	4.83	0.81	2.82	87	289	188
CA	146	876	511	752	9512	5132	6.85	1.14	4.00	87	289	188

Table 2: Estimate of the number of credentials for targets born in 1991 using a conservative set of parameters (C), a more liberal set of parameters (L) and the mean of the two (M). We furthermore show gross profits, cost per credential and the time needed (in minutes) to produce the credentials.

4 IRB Approval for studies

All studies presented in this paper were approved by the University's IRB board. The main experiment (in which students' online social network data was used to predict SSNs) was also vetted for FERPA compliance. Details of the approval (such as protocol approval numbers, and so forth) are omitted from this double-blinded submitted version, but can be made available to the editors and reviewers on request.

5 Communications with the SSA and other entities during the preparation of this manuscript

During our study, we contacted directly various departments at the SSA, as well as various industry experts at companies such as Choicepoint, TransUnion, and IDAnalytics, and several academic researchers in fields related to identity theft and Social Security numbers, to discuss issues related to the SSN issuance scheme and the feasibility of brute force verifications of the predicted SSNs. Copies of the private email exchanges can be made available to the editors and reviewers upon request.

References

B. Acohido and J. Swartz. *Zero Day Threat*. Sterling Publishing Co., Inc., New York, 2008.

A. Acquisti and R. Gross. Predicting Social Security numbers from public data. *Proceedings of the National Academy of Science*, 196 (27): 10975–10980, 2009.

N. Anderson. Monster resume information harvested by hacker; pishing attacks feared in future. ars technica, September 3 2007, 2007. http://arstechnica.com/news.ars/post/20070903-monster-resume-information-harvested-by-hacker-phishing-attacks-feared-in-future.html.

R. Anderson, R. Bohme, R. Clayton, and T. Moore. Security economics and the internal market. Report to the European Network and Information Security Agency, 2008.

C.Y. Cho, D. Babić, R. Shin, and D. Song. Inference and analysis of formal models of botnet command and control protocols. pages 426–440, 2010.

M. Cook. The lowdown on fraud rings. *Collections & Credit Risk*, 10, 2005.

E. Cooke, F. Jahanian, and D. Mcpherson. The zombie roundup: Understanding, detecting, and disrupting botnets. In *Workshop on Steps to Reducing Unwanted Traffic on the Internet (SRUTI)*, pages 39–44, June 2005.

Department of Homeland Security. I am an employer... how do I... use E-Verify? US Citizenship and Immigration Services, M-655, 2007. http://www.uscis.gov/files/nativedocuments/E4_english.pdf.

Department of Homeland Security. E-Verify user manual. M-574 E-Verify User Manual, 2008. http://www.uscis.gov/files/nativedocuments/E-Verify_Manual.pdf.

Electronic Privacy Information Center. Spotlight on surveillance: E-Verify system: DHS changes name, but problems remain for U.S. workers, 2007. http://epic.org/privacy/surveillance/spotlight/0707/default.html.

Financial Crimes Enforcement Network. Mortgage loan fraud: An industry assessment based upon suspicious activity report analysis, 2006. http://www.fincen.gov/mortgage_fraud112006.html.

FTC. Report to Congress under sections 318 and 319 of the Fair and Accurate Credit Transactions Act of 2003, 2004. http://www.ftc.gov/reports/facta/041209factarpt.pdf.

FTC. Staff summary of comments and information received regarding the private sector's use of Social Security numbers. FTC, Division of Privacy and Identity Protection, Bureau of Consumer Protection, 2007. http://www.ftc.gov/bcp/workshops/ssn/staffsummary.pdf.

FTC. Security in numbers: SSNs and ID theft: A FTC report providing recommendations on Social Security number use in the private sector. FTC Report, December 2008, 2008. http://www.ftc.gov/os/2008/12/P075414ssnreport.pdf.

GAO. Employment verification: Challenges exist in implementing a mandatory electronic verification system. Statement of Richard M. Stana, Director Homeland Security and Justice Issues. Testimony before the Subcommittee on Social Security, Committee on Ways and Means, House of Representatives. GAO-07-924T, 2007. http://www.gao.gov/new.items/d07924t.pdf.

E. Hendricks. Testimony of Evan Hendricks, Editor/Publisher Privacy Times www.privacytimes.com, before the House Committee on Financial Services Subcommittee on Financial Institutions & Consumer Credit, 2003. http://epic.org/privacy/preemption/hendricks6.12.03.pdf.

C. Herley and D. Florêncio. Protecting financial institutions from brute-force attacks. In *Proceedings of the Ifip Tc 11 23rd International Information Security Conference*, pages 681–685. Springer, 2008a.

C. Herley and D. Florêncio. A profitless endeavor: Phishing as tragedy of the commons. In *New Security Paradigms Workshop (NSPW)*, 2008b.

C.J. Hoofnagle. Putting identity theft on ice: Freezing credit reports to prevent lending to impostors. In A. Chander, L. Gelman, and M.J. Radin, editors, *Securing Privacy in the Internet Age*, pages 207–220. Stanford University Press, 2005.

C.J. Hoofnagle. Identity theft: Making the known unknowns known. *Harvard Journal of Law and Technology*, 21 (1): 98–122, 2007.

H.M. Howard. The negligent enablement of imposter fraud: A common-sense common law claim. *Duke LJ*, 54: 1263–1665, 2005.

ID Analytics. National report on identity fraud, 2003. http://www.idanalytics.com/. Document obtained from the company.

ID Analytics. U.S. national fraud ring analysis: Understanding behavioral patterns, 2005. http://www.idanalytics.com/. Document obtained from the company.

ID Analytics. National data breach analysis, 2006. http://www.idanalytics.com/. Document obtained from the company.

Internal Revenue Service. IRS warns of new e-mail and telephone scams using the IRS name; Advance payment scams starting, 2008. http://www.irs.gov/newsroom/article/0,,id=178061,00.html.

Irwin Financial Corporation. Comment to FTC's SSNs in the Private Sector, Project No. P075414, 2007. http://www.ftc.gov/os/comments/ssnprivatesector/531096-00310.pdf.

T.N. Jagatic, N.A. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Communications of the ACM*, 50 (10): 94–100, 2007.

M. Jakobsson and S. Myers. *Phishing and Counter-Measures*. John Wiley and Sons, 2006.

J. Leyden. Spammers open new front on social networking sites. The Register, May 14 2008, 2008a. http://www.theregister.co.uk/2008/05/14/social_network_spam/.

J. Leyden. Myspace wins record $230m judgement against spammers. The Register, May 14 2008, 2008b. http://www.theregister.co.uk/2008/05/14/myspace_spam_ruling/.

B. McCarty. Automated identity theft. *IEEE Security and Privacy*, 1 (5): 89–92, 2003.

L. Mitchell. New wrinkle in ID theft; thieves pair your SS number with their name, buy with credit, never get caught. Salt Lake Tribune June 6 2004, 2004.

T. Moore, R. Clayton, and R. Anderson. The economics of online crime. *The Journal of Economic Perspectives*, 23 (3): 3–20, 2009.

R. Mullins. The biggest cloud on the planet is owned by ... the crooks. *Network World*, March 22, 2010. http://www.networkworld.com/community/node/58829.

R. O'Harrow Jr. ID data conned from firm: Choicepoint case points to huge fraud. *Washington Post*, February 17: E01, 2005. http://www.washingtonpost.com/wp-dyn/articles/A30897-2005Feb16.html.

C. Papadimitriou. *Computational Complexity*. Addison-Wesley, 1994.

Samuelson Law, Technology & Public Policy Clinic. Comment to FTC's SSNs in the Private Sector, Project No. P075414, 2007. http://www.ftc.gov/os/comments/ssnprivatesector/531096-00295.pdf.

Selective Service System. Who must register, 2008. https://www.sss.gov/FSwho.htm.

AH Spencer. The Free Lunch: Arbitrage Profits Associated with Credit Cards. *Journal of Economic Issues*, 42 (1): 243, 2008.

SSA. Social Security number verification service (SSNVS) Handbook, 2007. http://www.ssa.gov/employer/ssnvs_handbk.htm.

SSA. Report fraud to the hotline, 2008. http://www.ssa.gov/oig/guidelin.htm.

Symantec. Symantec global internet security threat report, Trends for July-December 07, April 2008.

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

R. Thomas and J. Martin. The underground economy: Priceless. *;LOGIN:*, 31 (6): 7–16, 2006.

U.S. Census Bureau. Current population survey, general mobility by region, sex, and age: March 2000-2001, March 2001. http://www.census.gov/population/socdemo/migration/cps2001/tab01.pdf.

Websense. Google's CAPTCHA busted in recent spammer tactics. http://securitylabs.websense.com/content/Blogs/2919.aspx, 2008.

Westat. Findings of the web basic pilot evaluation. Report submitted to U.S. Department of Homeland Security, note = http://www.uscis.gov/files/article/WebBasicPilotRprtSept2007.pdf, note-accessed = Accessed on March 17, 2008, 2007.

L. Weston. How thieves steal their own identities. MSN Money, 2008. http://moneycentral.msn.com/content/Banking/FinancialPrivacy/P144890.asp.