AI Technologies to Defeat Identity Theft Vulnerabilities

Latanya Sweeney

Data Privacy Laboratory, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213-3890
latanya@cs.cmu.edu

Abstract

When a large number of citizens are at risk to identity theft, national security and economic prosperity are threatened. This work shows that thousands of Americans are at such risk, and introduces technology, named "Identity Angel," to help. Identity Angel's goal is to crawl through information available on the World Wide Web ("Web") and notify people for whom information, freely available on the Web, can be combined sufficiently to impersonate them in financial or credentialing transactions. This is an ambitious goal due to the various types of available information and the many inferences that relate disparate pieces of data. Therefore, the work presented herein focuses specifically on acquiring information sufficient to fraudulently acquire a new credit card using on-line resumes. An imposter needs to learn the {*name, Social Security Number, address, date of birth*} of a subject. Results show how resumes containing needed information can automatically be found and values harvested, and how many subjects removed such information from the Web once notified.

Introduction

Identity theft occurs when a person uses another person's personally-identifying information such as name, Social Security number ("SSN"), or credit card number, without permission to commit fraud or other crimes. People whose identities have been stolen can spend years –and lots of money— cleaning up the mess thieves have made of their credit record. Victims may lose job opportunities, be refused loans, education, housing or cars, or even get arrested for crimes they did not commit [FTC, 2002]. An editor at Consumer Reports examined credit card reports and found half those checked contained errors [2000]. The two reasons cited were being mistaken for another person with a similar name and fraud.

The Federal Trade Commission Report on Identity Theft [2002] shows rapid growth in victim complaints received at their clearinghouse. More than 86,000 cases were reported in 2001 and that grew to 162,000 cases in 2002. Nearly half of these involve credit card fraud. See Figure 1. Of credit card fraud, the report identifies more than half (or 26% of all thefts) as new accounts, making the acquisition of new credit cards, a major identity theft problem.

Incidents among younger adults were high. These adults are more likely to have resumes and facts about themselves posted on the World Wide Web ("Web"). They are also likely to have multiple residences in a short time period, making the issuance of a new credit card to a fraudulent address more difficult to determine. So, this work seeks to help this group by focusing specifically on the risk of new credit card fraud related to on-line resumes.

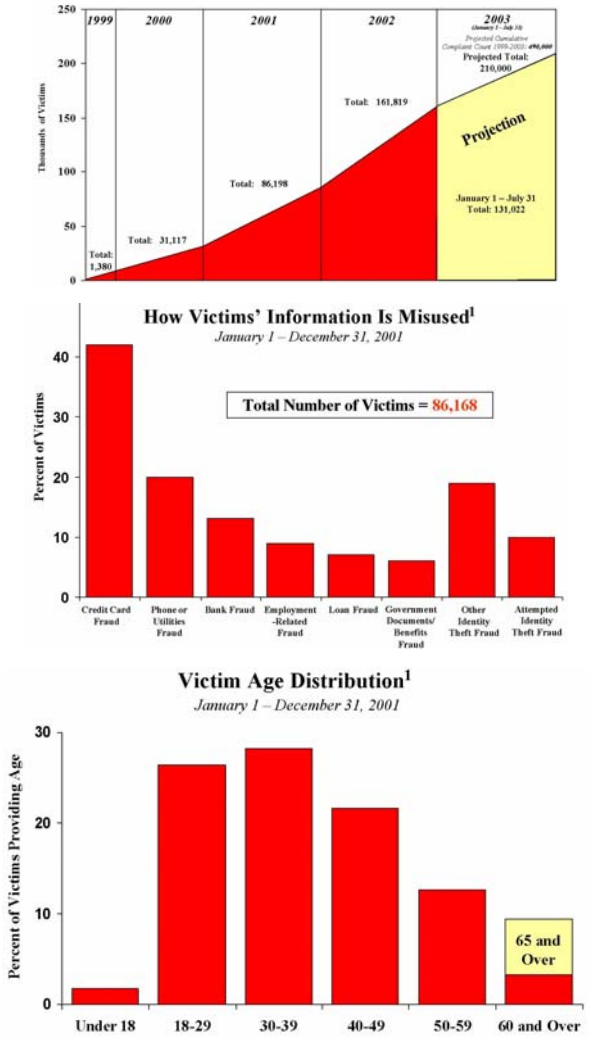

Figure 1. Information from the Federal Trade Commission Report on Identity Theft, based on victim complaints [2002].

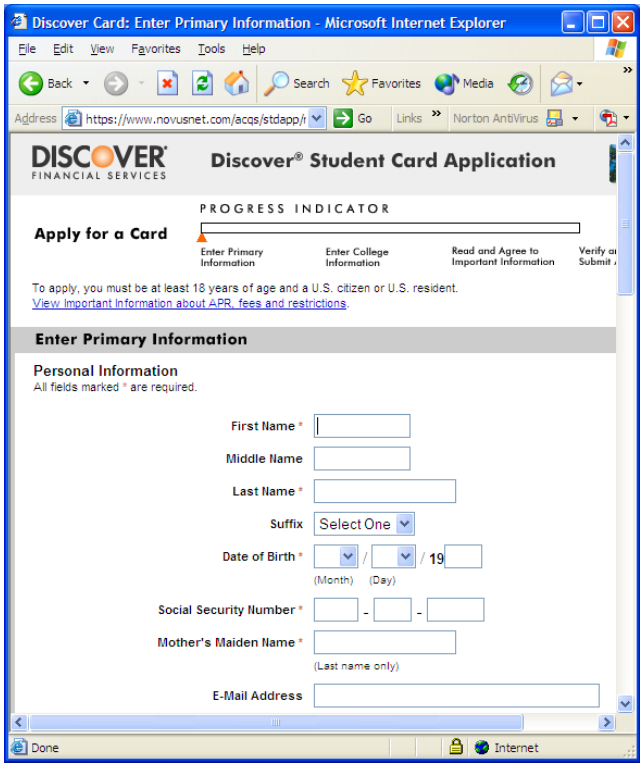

Figure 2. An on-line credit card application. Requested demographic information is shown and also includes phone and address (not shown). In student applications, the name of the college and expected year of graduation is also requested.

New Credit Card Fraud

In the acquisition of new credit cards, the person is only represented by the information provided on the application. The basic Information necessary for a credit card application is: {*Name*, *SSN*, *Address*, *Date of birth*, *Mother's maiden name*}; see Figure 2 for an example. How might an imposter gather the necessary information freely over the web? *Mother's maiden name* is used as a challenge question "after" the credit card is issued and not verified beforehand. The original *address* needs to be known, so a change of address can be included with the fraudulent application. Name searches on phone directories can often be used to find addresses. Several websites provide a *date of birth* ("*DOB*"), given a person's name (e.g., anybirthday.com). So, the most sensitive information is the SSN and its matching name.

Social Security Numbers

A key element to fraudulently acquiring a new credit card is the Social Security number. SSNs have evolved into national identifying numbers for individuals living and working in the United States. They are essential to identifying, recognizing, and authenticating people in health, financial, legal, and educational information. Some people might falsely believe that access to SSNs, while available within many financial, health, employment, and government institutions, are not publicly available.

The California-based Foundation for Taxpayer and Consumer Rights said for $26 each it was able to purchase the Social Security numbers and home addresses for Tenet, Ashcroft and other top Bush administration officials [Associated Press, 2003].

In contrast, the work reported herein addresses how to obtain SSNs for larger numbers of ordinary people using easily accessible and free on-line sources. In terms of risk, the less expensive sensitive information is to obtain and the greater the number of people having access to it, usually the greater the opportunity for abuse. Therefore, this work concerns SSNs available freely on the Web.

In 2003, the U.S. General Accounting Office identified SSN vulnerabilities as ripe for exploitation by terrorists [and other criminals], making SSN problems a serious concern to homeland security and a grave threat to the country's economic prosperity. These groups may gain funds for their activities while simultaneously causing havoc to the country's economy and citizens. The risks are real, yet poorly understood. The lack of scientific examination deprives society of possible benefits that may be realized by innovative technology. This work is proposed as one such technology.

Methods and Prior Work

Acquiring information sufficient to fraudulently obtain a new credit card using on-line resumes consists of locating on-line resumes, extracting information and emailing subjects. See Figure 3. The methods used to accomplish these tasks builds on the prior work described below.

In 2004, Sweeney introduced a system that locates on-line lists of names of people ("rosters"). Rosters are evasive to search engine retrieval because they do not lend themselves to keyword lookup. Using expressions such as "employees" or "students" returns hundred of pages, but finding the rosters among them previously required many hours of human inspection. Sweeney's approach ("filtered searching") executes a predicate function on each page retrieved from keyword searches to determine whether a page is an instance of the kind of page sought (e.g., a roster). The work reported herein uses filtered searching to locate on-line resumes.

In 1996, Sweeney used a system of entity detectors in a black board architecture to extract personally-identifying information from text files (e.g., letters and clinical notes). To date, the system continues to out-perform statistical and linguistic based approaches. The work reported herein also uses entity detectors, which in this case, are simple regular expressions, to identify instances of dates of birth, email addresses and SSNs appearing in resumes.

1	Locate on-line resumes (using Filtered Searching)
2	Extract sensitive values (using regular expressions)
3	Email subjects about their risks

Figure 3. Three processing steps required.

Results

Results show how on-line resumes containing needed information can automatically be found and values extracted and how some subjects protected such information from the Web once notified.

Richard Allen Brown. PO Box 782. Kayenta, AZ 86033. Home Telephone-520-697-3513. NAU Telephone-520-523-4099. **DOB: 03-10-77. SSN: 527-71 ...** dana.ucc.nau.edu/~rab39/RAB%20Resume.doc
…2843. **DOB: 10-10-48** New Britain, CT 06050-4010. F: (860) 832-3753. **SSN: 461-84-…** H: (203) 740-7255: (203) 561-8674. Education. Ph.D. www.math.ccsu.edu/vaden-goad/resume.htm
Scot Patrick Lytle. Home: (301)-249-5330 2116 Blaz Court School: (410)-455-1662 Upper Marlboro, MD 20772 **SSN: 578-90**-… userpages.umbc.edu/~slytle1/resume.html

Figure 4. Sample on-line resumes that include SSNs, Two of the resumes include dates of birth. All three include address and phone number. SSNs have been truncated for this writing but were fully available.

Materials

Experiments were based on: (1) FilteredSearch Java code; (2) two resulting resume databases; and, (3) entity detectors, as described below.

FilteredSearch. Java code that uses the Google API to perform a series of searches, pruning out duplicate pages.

Resume Databases. Using FilteredSearch on keywords {"resumes", "vitae"}, the first *n* distinct actual resumes containing SSNs were compiled into a database. *DBA* has 150 resumes from December 2003 and *DBB* has 75 resumes from December 2004 (excluding any in *DBA*).

Detectors. Regular expressions that identify ways of writing dates, SSNs and email addresses along with preceding heading such as "dob".

SSNwatch. To confirm whether a provided number was actually an SSN, the SSNwatch validation server was used to confirm that the number was likely to be valid (Sweeney, 2004).

Experiment 1: Finding Sensitive Resumes

Using FilteredSearch with a predicate function, which confirmed whether a retrieved page had format (using layout cues) and content (using headings) consistent with that of a resume or vitae, and included an SSN (using the appearance of 9 digits with and without an SSN heading and with and without dashes appearing between the digits. More than 2000 webpages of suspect resumes were filtered to produce *DBA* and similarly for *DBB*.

Based on exhaustive manual inspection of results and using the SSNwatch Validation Server, the following results were found. Of the 150 resumes in *DBA*, 140 (or 93%) had complete 9-digit SSNs. 10 resumes had partial, invalid, or some other country's SSN. All of the 75 resumes in *DBB* had 9 digit SSNs.

Experiment 2: Extracting Sensitive Information

Applying Detectors to *DBA*, and then manually inspecting each resume, provided the following results. All email addresses (113 of 113 or 100%) were found. The '@' and dot (.) notation worked well. All dates of birth (110 of 110 or 100%) were found, but some dates, which were not dates of birth were incorrectly reported as such; this happened in 20 cases (but only 7 where the proper DOB was not also found). SSN results were reported above.

In terms of combinations: 104 (or 69%) resumes had {*SSN*, *DOB*}; 105 (or 70%) had {*SSN*, *email*}, and 76 (or 51%) had {*SSN*, *DOB*, *email*}.

Experiment 3: Behavioral Impact

A single email message was sent to each of the 105 people in *DBA* having {*SSN*, *email*} alerting them to the risk. A year later, 102 (or 68% of all of *DBA*) no longer had the information available. In *DBB*, 46 were notified, and within a month, 42 (or 55% of all of *DBB*) no longer had the information publicly available.

Discussion

Imagine a benevolent program that emails people for whom information, freely available on the Web can be combined sufficiently to impersonate them in financial transactions. This is the ambitious goal of "Identity Angel." The work reported herein focused on a sub-problem of Identity Angel –namely, acquiring information sufficient to fraudulently acquire a new credit card using on-line resumes. This work demonstrated the viability of this approach.

Acknowledgements

Thanks to Sylvia Barrett, Kishore Madhava, Yea-Wen Yang and Nicholas Lynn for assistance.

References

Consumer Reports, July 2000.

Social Security numbers sold on Web. Associated Press, 8/2003.

Sweeney, L. Finding Lists of People on the Web. *ACM Computers and Society*, 34 (1) April 2004.

Sweeney, L. Replacing Personally-Identifying Information in Medical Records, the Scrub System. *Proceedings, Journal of the American Medical Informatics Association*. 1996.

Sweeney, L. *SSNwatch Validation Server*, 2004 <privacy.cs.cmu.edu/dataprivacy/projects/ssnwatch/index.html>.

United States Federal Trade Commission, *Report on Identity Theft, Victim Complaint Data: Figures and Trends January-December 2001*, Federal Printing Office, Washington, DC: 2002.

United States General Accounting Office, *Improved SSN Verification and Exchange of States' Driver Records Would Enhance Identity Verification*, Federal Printing Office, Washington, DC: 2003.