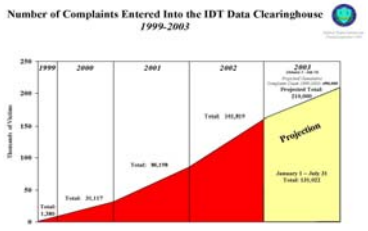
Carnegie Mellon
DATA PRIVACY LAB

Identity Angel

AI Technologies to Defeat Identity Theft Vulnerabilities

Identity Angel's goal is to scan the Web determining whether there is sufficient publicly-available information on an individual to fraudulently represent the person in financial and credentialing transactions. The experiments reported here focus specifically on acquiring information sufficient to fraudulently acquire a new credit card using freely available on-line resumes.

An imposter needs to learn the {name, Social Security Number ("SSN"), 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.



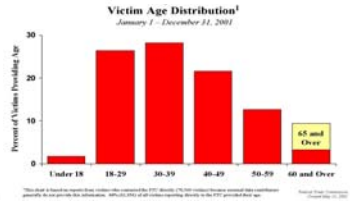
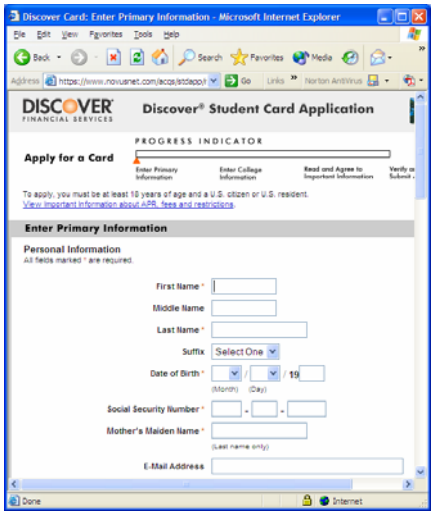


Figure 1. The Federal Trade Commission Report on Identity Theft [2002] shows rapid growth in victim complaints. Nearly half involve credit card fraud. More than half of these (or 26% of all thefts) involve new accounts, making the acquisition of new credit cards, a major identity theft problem.

Figure 2. The basic Information necessary for a credit card application is: {*Name*, *SSN*, *Address*, *Date of birth*, *Mother's maiden name*}. 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. So, the most sensitive information is the *SSN* and its matching *name*. 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.



Filtered Search

In 2004, Sweeney introduced "filtered searching" as a means to locate on-line rosters of names. Rosters, like resumes containing SSNs, do not lend themselves to direct keyword search. Expressions like "resume" and "SSN" returns hundred of pages, but finding the actual resumes among them previously required hours of human inspection. Filtered search does automatic retrieval and then executes a predicate function to determine whether the page is a resume of interest.

SSNwatch

To confirm whether a provided number is an actual SSN, the SSNwatch Validation Server was used (Sweeney, 2004). See privacy.cs.cmu.edu/dataprivacy/projects/ssnwatch/index.html.

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

Figure 3. Three processing steps.

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.

FilteredSearch was used 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.

1000 resume hits on Google, revealed 150 resumes, of which 140 (or 93%) had complete 9-digit SSNs.

All email addresses (113 of 113 or 100%) were found.
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.

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

A single email message was sent to each of the 105 people having {SSN, email} alerting them to the risk.
Within a month, 42 (or 55%) no longer had the information publicly available.
A year later, 71 (or 68%) no longer had the information available.

Latanya Sweeney
http://privacy.cs.cmu.edu/dataprivacy/projects/idangel/index.html