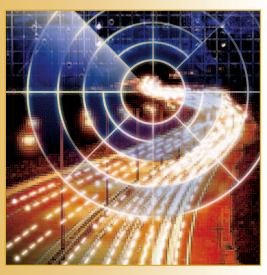

Editors: Elisa Bertino • *bertino@cerias.purdue.edu*
Steve Ruth • *ruth@gmu.edu*

Protecting Job Seekers from Identity Theft

In search of a job after graduating from college, Meg Kemp posts her resume online. Norton Steuben, a retired law professor, hasn't looked for employment in more than 35 years and rarely uses the Internet, yet his law school maintains his curriculum vita online. Although such activity might seem innocuous, information in both these online resumes initially put Meg and Norton in danger of becoming identity-theft victims. Fortunately, they received some unexpected protection in the form of a computer program called Identity Angel, which this article describes in more detail.

Latanya Sweeney
Carnegie Mellon University

Identity theft is a growing problem, and personal information included in online resumes and vitae can increase the risk of being a victim. In many cases, people in pursuit of employment willingly provide resume information; in other cases, organizations post employee and affiliate vitae, sometimes without their knowledge or consent.

Of course, nobody's proposing that resumes and vitae shouldn't be posted on the Internet. Instead, people should take care in selecting which information to include — online resumes and vitae shouldn't contain social security numbers (SSNs), for example. They should substitute age for date of birth (although publication of this number should be avoided whenever possible) and include home addresses only if they're absolutely necessary. Once a resume appears on the Internet, it's likely to be eternally public, thanks to search engine caches and Internet archiving organizations.

The best remedy is to never post sensitive information online in the first place, but if your resume information is already out there, you must do whatever you can to limit your exposure. Wouldn't it be helpful if someone could notify you and make you aware of the risk and how to mitigate it? Informed subjects are likely to remove sensitive information, so personalized notification and education provide an effective means of achieving this outcome. But how can we find these people? And more importantly, how do we notify them? This task isn't trivial — careerbuilder.com, for example, claims to host 14 million online resumes.

Technology can help. Imagine a benevolent program that can crawl through freely available information on the Internet and email people for whom uncovered

 Published by the IEEE Computer Society 1089-7801/06/$20.00

facts can be combined to impersonate them in financial or credentialing transactions. This is the ambitious goal of the Identity Angel Project at Carnegie Mellon University's Data Privacy Laboratory (http://privacy.cs.cmu.edu/dataprivacy/projects/idangel/index.html). Identity Angel locates resumes that contain sufficient information to fraudulently acquire a new credit card and then notifies the subjects of these resumes by email, encouraging them to remove sensitive information. This article describes the risks associated with freely available resume information and how Identity Angel successfully mitigates those risks.

The Nature of Identity-Theft Problems

Identity theft occurs when one person uses another person's identifying information without permission to commit fraud or other crimes (see the "The Identity Theft Problem" sidebar). The US Federal Trade Commission's (FTC's) report on identity theft[1] shows a rapid growth in victim complaints. Victims reported more than 86,000 incidents in 2001; this number grew to 162,000 cases in 2002, and rose to 246,570 in 2004. More than a quarter (28 percent) of reported complaints involved credit-card fraud. Of credit-card fraud complaints, the report identified that approximately half (or 17 percent of all thefts) involved new accounts, making the acquisition of new credit cards the major identity-theft problem, as Figure 1 indicates.

Most incidents (roughly 29 percent) occurred among younger adults, who tend to have more resumes and facts about themselves posted online. They also tend 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.

Fortunately, US courts have maintained that a victim isn't liable for charges made by a credit card issued to someone else who forged the victim's name on the application — provided the victim knew nothing about the credit card.[2] Although the victim doesn't have to pay for the goods and services purchased on a fraudulently acquired card, the expense is paid by everyone else in the system through higher interest rates and aggressive fee structures.

The worst part for the victim is in trying to clean up his or her credit report. People whose identities have been stolen literally spend years — and lots of money — trying to clean up the mess. Moreover, identity thieves tend to acquire several cards in the victim's name, which means that each of them must be corrected separately. The existence of these cards can go undetected by the victim for years, making correction even more difficult once discovered. For several years later, victims have reported lost job opportunities and loan refusals for education, housing, and cars; some have even been arrested for crimes they didn't commit.[3] Several online logs, such as Identity Theft Spy (www.identitytheftspy.com), archive cases that appear in the press.

To stem the tide, the FTC encourages people to review their credit reports annually (www.consumer.gov/idtheft/con_minimize.htm). This is a good practice overall, but it carries a caveat. If an unknown credit card appears on a credit report, the victim must immediately begin correction and notification procedures or else risk being forever responsible for the debt. Once the victim receives the credit report, he or she is considered to have been officially "notified." Inaction on the victim's part can be interpreted as authorizing the card's use. The US National Association of State Public Interest Research Groups surveyed 197 people and found half of their reports contained errors; the two reasons cited were being mistaken for another person with a similar name and fraud.[4]

The Identity Theft Problem

As more people use the Internet for personal and business matters, it isn't surprising that criminals have adapted their methods to steal from people online. Although this article concerns information people post about themselves in online resumes, email-related attacks are also common. Criminals attempt to mask their fraudulent intent by luring readers to perform actions that expose their sensitive information, monetary funds, or identity information. Two common examples of fraudulent email include

- *Phishing*, which is the act of sending an email message impersonating a respected organization in an attempt to get the reader to click on the provided link and give personal information. The reader believes he or she is communicating with the respected organization, when in reality, the criminal is harvesting account, password, and other personal information the reader provides.
- *Advance fee scam*, which is an email message that fraudulently offers the opportunity for the reader to receive a large amount of funds. Examples include winning a lottery or helping secure funds from an isolated bank account. In these cases, the reader typically pledges a smaller amount of money or provides personal account information as collateral.

For more information on these two forms of attack, visit http://antiphishing.org, www.secretservice.gov/alert419.shtml, and privacy.cs.cmu.edu/dataprivacy/projects/scamspam/.

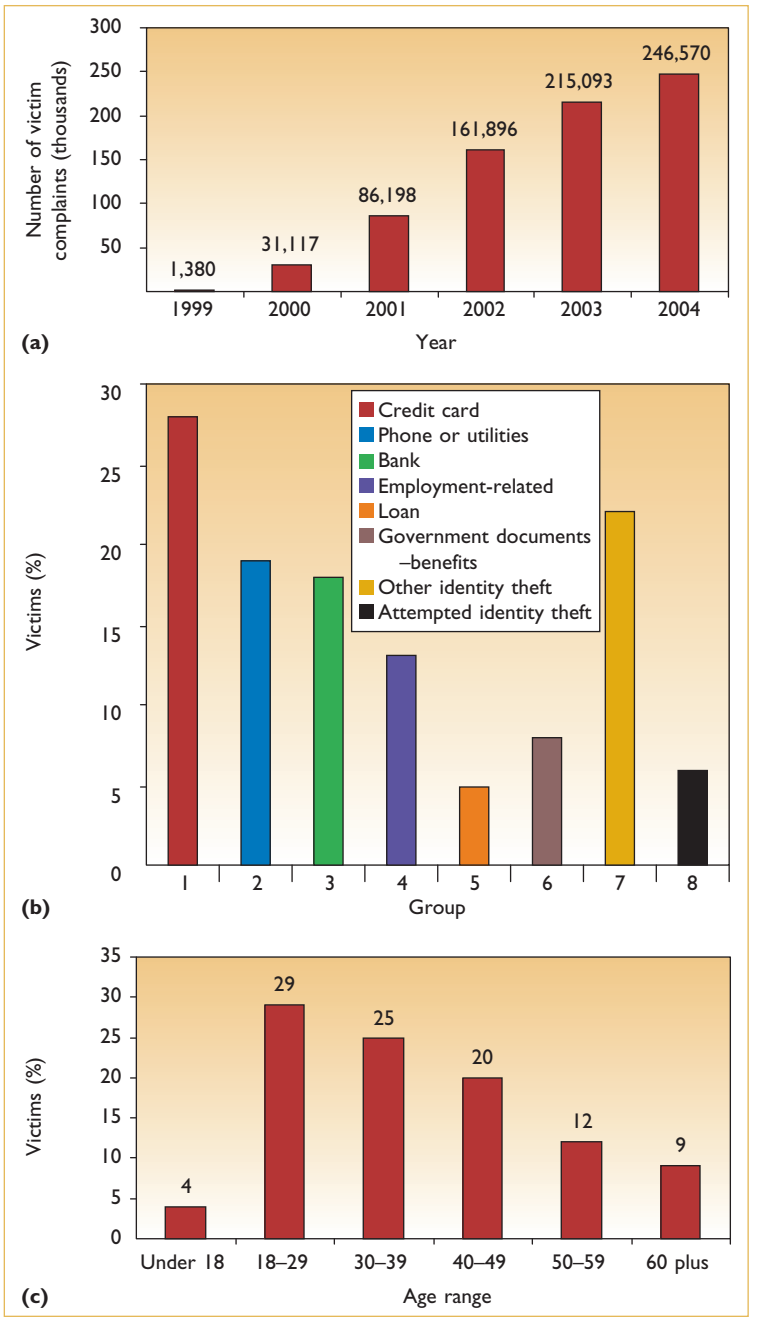

Figure 1. Identity theft. Information from the US Federal Trade Commission's report on identity theft indicates (a) a rise in the number of complaints, (b) a large percentage of victims experiencing credit-card fraud, and (c) the largest age demographic.

Applying for a New Credit Card

When you apply for a new credit card, you're only represented by the information provided on the application itself. The basic data requested on a credit-card application is your name, SSN, address, date of birth, and mother's maiden name. Figure 2 shows an example.

So how could an imposter gather the necessary information about you, freely, over the Web? Your mother's maiden name is used as a challenge question after the credit card is issued (it isn't verified beforehand), thus any name will suffice in most cases. The original address must be known, so a change of address can be included with the fraudulent application; name searches in phone directories can bring up most addresses. Several Web sites provide a date of birth, given a person's name (for example, anybirthdate.com). So, the most sensitive information an imposter needs to get is your SSN.

SSNs have evolved from being internal account numbers for the Social Security Administration into national identifying numbers for individuals living and working in the US today. They're essential for identifying, recognizing, and authenticating people in health, financial, legal, and educational contexts. Amazingly, some people still believe that SSNs aren't publicly available. In 2003, the California-based Foundation for Taxpayer and Consumer Rights paid US$26 each for the SSNs and home addresses of some of President George W. Bush's top officials.[5] Also in 2003, the US General Accounting Office[6] identified SSN vulnerabilities as ripe for terrorist exploitation, making such vulnerabilities a serious concern to homeland security and a grave threat to the country's economic prosperity.

Methods for Mining Resume Information

In 2004, I introduced a system that locates online lists of people's names, or *rosters*.[7] These rosters evade search engine retrieval because they don't lend themselves to keyword lookup. Using expressions such as "employees" or "students" returns hundred of pages, but finding the rosters among them takes many hours of human inspection. My approach, which I call *filtered searching*, executes a simple predicate function on each page retrieved from keyword searches to determine whether a given page is an instance of the kind of page sought.

Identity Angel uses filtered searching to locate online resumes. Entering the word "resume" in a search engine yields more than 100 million Web pages about resume writing and resume submissions in addition to resumes themselves. Filtered searching confirms whether a given Web page has format and content consistent with a resume or vita. One way to identify a Web page that contains a resume is to look at its layout – resumes have a structured layout with bulleted items grouped by headings. In comparison, Web pages about resume

writing usually have paragraphs of complete sentences. Filtered searching exploits the appearance of a Web page's layout and item headings to decide whether a given page is likely to be a resume.

In 1996, I used a system of *entity detectors* in a blackboard architecture to extract personally identifying information from text files (such as letters and clinical notes).[8] One detector identified appearances of first names, another detected addresses, another identified dates, and so on. Each detector was an expert at identifying occurrences of its assigned entity. Precedence was used to resolve differences among detectors and to infer other information. No attempt was made to try to understand the text. To date, the system continues to out-perform statistical and linguistic-based approaches that use grammar and syntax processing of sentences.

My Identity Angel approach also uses entity detectors to identify instances of birth dates, email addresses, and nine-digit SSNs appearing in resumes. This is particularly important with resumes, which tend to have phrases, not complete sentences. These detectors exploit the ways in which we traditionally write dates, email addresses, and SSNs: the "@" and dot (".") notation in email addresses, for example, is a writing convention unique to email addresses. Similarly, people usually write SSNs either as nine continuous digits or three digits, dash ("-"), two digits, dash, followed by four digits. In resumes, key phrases such as "date of birth," "SSN," or "social security number" appear adjacent to their related values. Identity Angel exploits such writing conventions to harvest sensitive information from resumes.

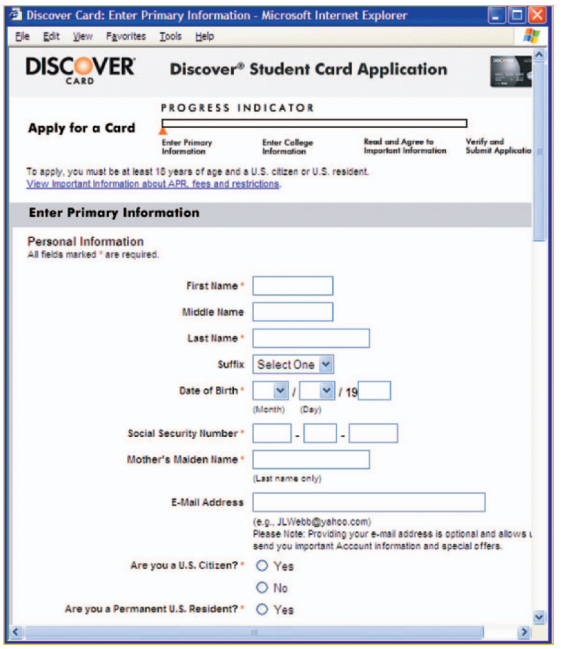

Figure 2. An online credit-card application. Requested demographic information includes the applicant's name, date of birth, social security number, mother's maiden name, citizenship status, and phone number and address (not shown). In student applications, the name of the college and expected year of graduation is also requested (https://www.discovercard.com/cardmembersvcs/discovercard/apply-for-a-card/primaryinformation).

Experimental Results

We recently used Identity Angel in some experiments to locate publicly-available resumes that had information sufficient to fraudulently acquire a new credit card, and to communicate with the subjects of this fraud to see if email notification would lead them to remove that information.

We wrote a Java program that used filtered search and the Google API to identify resumes, and entity detectors for SSNs, dates of birth, and email addresses to extract information from found resumes. We conducted the first experiment in December 2003 and the second in December 2004. We started by processing the first 2,000 Web pages (of millions) retrieved from Google searches on "resume," "vitae," and "SSN." In 2003, we found 150 resumes among the first 2,000 Web pages found; in 2004, we found only 75 resumes among the first 2,000 pages, excluding any resumes that appeared in the 2003 experiment. For convenience, we call the 150 resumes from the 2003 experiment "DBA" and the 75 resumes from 2004 "DBB." To confirm whether a value provided by the SSN was actually an SSN, we used the SSNwatch Validation Server (http://privacy.cs.cmu.edu/dataprivacy/projects/ssnwatch/index.html).

Sensitivity of Resumes

Figure 3 shows samples of some of the resumes we found. Although the information provided is truthful, some information has been omitted for privacy reasons. As we learned earlier, the SSN is the most sensitive value. Of the 150 resumes in DBA, 140 (or 93 percent) had complete nine-digit SSNs, whereas 10 had partial, invalid, or some other country's number. All of the 75 resumes in DBB had nine-digit SSNs.

Extracting Sensitive Information

We manually reviewed each of the resumes in DBA

Richard *****. Kayenta, AZ 86033.
Home Telephone-520-697-****. NAU Telephone-520-523-****.
DOB: 03-10-77. SSN: 527-71-****.
http://dana.ucc.nau.edu/~rab39/RAB%20Resume.doc

…2843. DOB: 10-10-48 New Britain, CT 06050-4010. F: (860) 832-****.
SSN: 461-84-**** H: (203) 740-**** : W: (203) 561-****. Education. Ph.D.
www.math.ccsu.edu/vaden-goad/resume.htm

Scott *****. Home: (301)-249-****. School: (410)-455-****. Upper Marlboro, MD 20772
SSN: 578-90-**** http://umbc.edu/~slytle1/resume.html

Figure 3. Sample online resumes. Two of the resumes include dates of birth, and all three include addresses and phone numbers. Social security numbers are truncated here, but were fully available online.

and recorded occurrences of dates of birth (DOBs), email addresses, and SSNs. In summary, 104 (or 69 percent) of the resumes had {SSN, DOB}; 105 (or 70 percent) had {SSN, email}, and 76 (or 51 percent) had {SSN, DOB, email}.

We then compared these results to the values extracted by the detectors. In DBA, the detectors found all the email addresses (113 of 113, or 100 percent). The "@" and dot (.) notation also worked well. The detectors also found all the DOBs (110 of 110, or 100 percent), but some dates were incorrectly reported as such; this happened in 20 cases (but only seven in which the proper DOB wasn't also found).

Behavioral Impact

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

Although other countries have national identification numbers, they don't face the same risks as US citizens, who use their SSNs in many different situations. Canada, for example, has a number similar to the US SSN, but by Canadian law, this number can't be used as part of credit-card issuance. People don't always know the risk they place themselves in when they put information online, but when notified of actual risks and harm, they tend to take corrective action. Of course, education and notification have always been tools in public policy, but in this case, technology specifically targets those involved. Rather than simply disseminating general promotions and announcements, the ability to provide targeted intervention is a clear benefit of the Internet.

As a next step, we intend to operate Identity Angel on a wide-scale basis in 2006, so it can notify thousands of people whose online resumes make them vulnerable to identity theft. We also intend to establish a Web site for people to submit resumes for a privacy review.

Acknowledgments

Thanks to Elisa Bertino for the opportunity to present this article and to Sylvia Barrett, Kishore Madhava, Yea-Wen Yang, and Nicholas Lynn for data assistance. This work was conducted by volunteer service from the Data Privacy Laboratory in Carnegie Mellon University's School of Computer Science. Government and/or business sponsors are welcomed.

References

1. US Federal Trade Commission, "Report on Identity Theft, Victim Complaint Data: Figures and Trends January-December 2004," Federal Printing Office, 2005.
2. D. Szwak, "Understanding Credit Cards, Credit Reports and Fraud," *The 'Lectric Law Library*, Jan. 2006; www.lectlaw.com/files/ban16.htm.
3. T. Zeller, "Waking up to Recurring ID Nightmares," *The New York Times*, 9 Jan. 2006; www.sftnj.com/news/pdf/nytimes.pdf.
4. US Nat'l Assoc. of Public Interest Research Groups, *Mistakes Do Happen: A Look at Errors in Consumer Credit Reports*, June 2004; http://uspirg.org/reports/ MistakesDoHappen2004.pdf.
5. K. Edwards, "Social Security Numbers Sold on Web," Associated Press, Aug. 2003; www.cnn.com/2003/TECH/internet/08/28/privacy.concerns.ap/index.html.
6. US General Accounting Office, "Improved SSN Verification and Exchange of States' Driver Records Would Enhance Identity Verification," Federal Printing Office, 2003.
7. L. Sweeney, "Finding Lists of People on the Web," *ACM Computers and Soc.*, vol. 34, no. 1, 2004; http://privacy.cs.cmu.edu/dataprivacy/projects/rosterfinder/index.html.
8. L. Sweeney, "Replacing Personally-Identifying Information in Medical Records: The Scrub System," *Proc. J. Am. Medical Informatics Assoc.* 1996; http://privacy.cs.cmu.edu/people/sweeney/scrub.html.

Latanya Sweeney is an associate professor of computer science, technology, and policy at Carnegie Mellon University. Her research interests include developing privacy-enhancing technologies. Sweeney has a PhD in computer science from the Massachusetts Institute of Technology. Contact her at latanya@privacy.cs.cmu.edu; http://privacy.cs.cmu.edu/people/Sweeney/index.html.