The Internet Hunt Revisited: Personal Information Accessible via the Web

Kay Connelly, Tom Jagatic, Ashraf Khalil, Yong Liu, Katie A. Siek and Sid Stamm
Indiana University

1 Introduction

In June 1993, Rick Gates posted an unusual Internet Hunt[4] to his monthly contest in usenet (alt.internet.services): he presented a simple email address and asked his contest participants to find out what they could about the owner of the email address.

The participants found 148 "separate pieces" of information including the subject's employer, his addresses, and his fiancee's name[5]. This immense availability of information was surprising especially considering the subject's employer: the Central Intelligence Agency.

We have entered the age of identify theft, online romance and virtual communities. Given that all but the most extreme luddites have used the Internet at one time, it is important to know how much can be discerned about an individual by simply surfing the web. Can a person stalk you or steal your identity from half way around the world without leaving their desk?

We investigate the Internet in 2004 to see how readily available information is on the web, classify the types of information we find and the sources we use and discuss the implications in today's growing web-centered world.

1.1 Rules of Engagement

For the purpose of our search, we limited our methods to what could be done legally on the web. As we scoured for information, we kept in mind the following rules:

- We shall only use the web. No "real-world" visits or phone calls.
- We will not make contact with the subjects who are the targets of our hunt.
- We shall not go to a library or any physical record archive.
- We shall not break the law.

Using these criteria, we searched the web for whatever information we could find about our chosen subjects.

1.2 Overview

In Section 2, we describe how and why we chose our targets and discuss the tools and techniques we used. We present a classification for information and web sources in Section 3. In Section 4, we describe the results of our search: what we found and where we found it. We conclude in Section 5.

2 The Setup

In this section, we discuss the targets of our hunt, the tools we use and our hunting methods.

2.1 Choosing Targets

When choosing our targets, we had several criteria. To ensure a web presence, each target should have at least one active email address and some other type of web activity (e.g. web page or active poster to an online forum). To minimize false positives, we chose targets who did not have common names. To demonstrate the potential serious nature of (or the lack thereof) web privacy, one target was to be active oversees in the US military. We wanted at least one of our targets to maintain a weblog (diary pages with entries presented in reverse-chronological order also known as a *blog*), since many people reveal quite personal information on them. Also, we wanted one target to be over 40 and the other to be younger than 25 so we could examine both a generation that grew up with the Internet and another that did not.

We chose two targets with very different profiles for our search. Our first target, Alice[1], is in academia and over 40. She has a web page through her work and has been using the Internet for several years. Our second target, Bob is in the US military reserves and was serving in the Middle East at the time of our search. When we chose Bob, we knew he was under 25 and had one blog. Later we discovered that he maintained a second blog.

2.2 Resources and Tools

Logging and Data Archiving In order to keep track of the time spent scouring the web, to store important documents for later analysis, and to share the information that was found with other team members, we developed an online logging and data archiving system. The logging and data archiving system consists of a web login interface, data input/display forms, and a back end database. All of the searching activities of the hunt team were performed through this system. During each searching activity, the system kept records of the hunter's identity, searching time (between login and logout), information about the documents found by the hunter (document title, source, credibility, significance, etc.), and facts about the subjects drawn from these documents. The data archiving system also kept local copies of some important documents with the subjects' private information on it (e.g. pictures, background reports, etc.). The team made use of these copies in later discussions and analysis.

A major goal of the logging system was to be able to draw a timeline of facts found about each target in the Internet Hunt process. This helped us determine when each piece of information was spotted, how much effort was spent on it and if it resulted in the ability to find more information.

The system also helped us validate the facts when we had contradicting information because the hunt team could easily double-check the sources that were in the database relating to a particular piece of information.

Webwhacker When we needed to keep a local copy of part of a website for later analysis, we used a website downloading tool called Webwhacker. Webwhacker helped us to quickly create an archive of the site on the local hard drive.

Onion Router During the hunt, we did not want our targets (or anyone else) to be aware of our identity. We made use of onion routers (OR) in some of our web searching when we had to visit a site frequently (e.g. to check on Bob's blog). ORs work like anonymizing proxies. In addition to making our identity invisible to the visited sites, they also hide the content of the communication from eavesdroppers up to the point where the traffic leaves the OR network.

[1] Names have been changed to protect the innocent.

2.3 Procedure

The Hunters Our research team for the Internet Hunt consisted of five hunters. The hunters were divided into three groups - Alice research group, Bob research Group, and pay site group. The Alice and Bob research groups only used web sites and search engines that did not require a fee. The pay site group researched both Alice and Bob on web sites that required a fee for searching or obtaining information. There was a lead hunter who coordinated the work of the three groups and facilitated sharing of information, like web sites that could be used to search for both targets.

Web Searching The major task of the Internet Hunt was web searching. Search engines such as Google[1] and Yahoo[2] were the major tools during this process. Some professional investigating web sites were also used during the searching. Hunters recorded all of their findings via the data archiving system. They evaluated each relevant document found on the web, assigned a credibility value and significance value to it, and kept a local copy of the document. The documents were in various forms (PDF, HTML, and DOC). After a document was found, one or more facts were deduced from it and submitted to the data archiving system.

Meetings The team met each week to discuss searching results and set goals for the next week. Hunters took this opportunity to exchange their thoughts about their searching methods and techniques. Such discussion helped to evaluate hunters' searching activities in the previous week. It also improved the whole team's searching efficiency.

Milling Over the Information After the Web searching was done, the team worked together to mill over all the information and their related documents found in the previous phase. At this time, we chained information together to determine new information. For example, Figure 1 shows that a background check gave us the name of an older man who lived at most of the same residences as Bob. We suspected this was Bob's father, but could not be sure. Bob mentioned he started a business with his father in his blog. We looked up the business but could not find any names. We then looked up the business' main product in a patent database and found Bob's name with the co-business owner — his father. Since the name matched the name from the background check, we were able to chain the information together to create a new piece of information — the name of Bob's father.

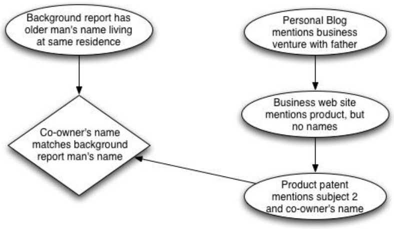
Background report has older man's name living at same residence
Personal Blog mentions business venture with father
Co-owner's name matches background report man's name
Business web site mentions product, but no names
Product patent mentions subject 2 and co-owner's name

Figure 1: Chaining for Bob's father's name

In addition, a major activity timeline (including home address changes) was created for each subject. The information gathered about each subject was categorized and counted. The time and expense we spent on the hunt were calculated as well.

3 Classifications

This section introduces the way in which we will classify the information and source web sites. The information is categorized according to the degree of control the target has over the information, the type of web site which provided us with the information and the type of information.

3.1 Control over Information

The information we gathered can be categorized into voluntary and involuntary information. By voluntary, we mean the information that is published with the full awareness and approval of our target, regardless of the information publisher. Involuntary information is that which is published online without the approval, and possibly, the knowledge of our target. Moreover, the person does not have any control over involuntary information. For example, Alice's homepage is voluntary, but her credit and criminal reports are not.

3.2 Types of Web Sources

There are many online sources that provide personal information. From the gathered information, we identified the following types of sources:

Personal A personal website such as a homepage or a blog is often the most accessible and intuitive source of information. The type and amount of personal information posted on such sites varies significantly, but often includes information about the person such as how to contact them, names of friends, hobbies, daily activities, weakness, strengths, moods, and jobs.

Official Official websites are those that are published by official organizations such as government agencies, universities, and corporations. Such organizations often publish personal information about the people affiliated with them but it is often common knowledge such as name, title, and contact information.

Media Media sites such as newspapers and organization periodicals contain articles which may quote people or describe a newsworthy activity in which the person is involved.

Contributions People may contribute to some public website, newsgroup, discussion list, or online group that takes the form of advice, opinion, feedback, or announcements. Such contributions can provide valuable information about the contributor. Web archiving magnifies the effect of such contributions and makes them everlasting. Information contributed in the past may reflect opinions that have since changed and may inaccurately and even negatively characterize a person.

Pay Sites There are several companies that provide background and criminal checks for a fee. Only one pay site explained that they got their information from a third party that pays people to visit record buildings and input data manually. Such sites will provide aliases which the person has used, current and past addresses, names of neighbors and any tickets, warrants or arrests.

3.3 Types of Information

In order to better understand the kinds of information that we were able to obtain, we classify the types of information as follows:

Education includes information about the degrees a person holds and the schools and universities attended by that person.

Career is all the relevant information about career history. This includes the name of the employers, the career's title, dates of employment, and salary.

Military Service information includes all the relevant information about the military service of that person, such as the service dates, rank, unit, and responsibilities.

Location information includes details about a person's residential history, including addresses, dates of residence, and neighbors.

Family information includes information about both the immediate and extended family. Personal information about family members is often directly correlated to the subject's personal information.

Interests and Social Life include information about activities, friends, hobbies, recreational activities, volunteering and other social interactions.

Political information includes the political orientation, the degree of involvement in the political system, voting history, monetary donations to political parties, and involvement in political organizations.

4 What We Found

During our search, we found 120 different pieces of information about Alice and 253 about Bob. We spent approximately 50 man-hours scouring the web[2], which is an average of 8 minutes per piece of information. We spent a little over $400 on pay sites.[3]

In this section, we first describe the amount of information we found on each subject and classify the information and web sites that we used. We then give a more qualitative discussion of the web sources and information.

Category	Alice	Bob
Education	3	17
Career	6	4
Service	0	28
Location	38	25
Family	62	62
Interests and Social Life	1	58
Political	0	5
Other	10	54
Total Information:	120	253

Table 1: Information Classified by Type

4.1 Classification of Information

Table 1 shows the number of pieces of information we found, categorized by type as described in Section 3.3.

Topping the list was information about our targets' families (62 for both Alice and Bob), including names, addresses and birthdays of immediate family members, Bob's mother's maiden name, the purchase price of Bob's parent's home, and the salary, college grades and employer of one of Alice's children. We also found several photos of the targets' families, including a 1958 reunion picture of Alice's family.

For both targets, we found a large amount of information about their current and past locations (38 for Alice and 25 for Bob) including a history of past addresses from which we could construct a detailed timeline for each target. In addition to their current home addresses and phone numbers, we obtained satelite photos of their neighborhoods indicating that Bob lives on a culdesac and Alice lives in a heavily wooded subdivision. We also found information about their current and past neighbors.

We found 58 disctinct pieces of information about Bob's personal life including organizations he belongs to, awards he won in college, movies and books that he likes, friend's names and stories about them and magazines he reads. The majority of the personal information was found from his multiple blogs.

Not surprisingly, we found no information about Alice's military service, as she never served. We did find 28 pieces of information relating to Bob's service, including his rank, distinctions, job duties, Company Commander's name and details about where he was currently stationed in the Middle East.

Under education, in addition to the names of high schools and colleges for both targets, we found Bob's college GPA and that he took AP European history while in high school.

[2] Because of some searching session which were never logged out, we do not have a precise time.

[3] If we had known in advance the quality of the paysites (Section 4.2), we would have spent less than $200 to obtain the same information.

	Alice		**Bob**	
Category	**# Sources**	**# Facts**	**# Sources**	**# Facts**
Personal	5	77	3	119
Official	2	2	1	9
Media	0	0	3	13
Contribution	0	0	2	34
Pay Sites	3	75	3	81
Total:	10	154	12	259

Table 2: Information Classified by Source

Miscellaneous information listed as other in the table includes Alice's birthdate, public PGP key, how much she bought and sold her house for (and thus the profit made) and several pictures. For Bob, we found his actual signature from a scanned in document, multiple email addresses and websites, his sexual orientation, religion, astronomical sign, childhood nickname and many pictures.

Table 2 shows the categorization of information by web source, as described in Section 3.2. For Alice, 77 facts were obtained from a total of 5 personal sites, including a site maintained by Alice's husband and one maintained by her child. Two official sites provided only two facts. Three pay sites, however generated 75 facts, some of which overlapped with the information we found on the personal sites. We did not find any information about Alice from media sites or public contributions.

For Bob, a total of 119 pieces of information were found from his 3 personal sites. 9 facts were obtained from one official web site, 34 from contributions he made to two public forums, 81 facts from three pay sites, and 13 from three different media sites, including his college's student newspaper.

Subject	**# Voluntary**	**# Involuntary**
Alice	78	49
Bob	153	107

Table 3: Information Classified by Control

Finally, Table 3 shows the classification of the information we found based on if it was voluntarily disclosed or not as described in Section 3.1. For both Alice and Bob, around 40% of the information we found was not voluntarily disclosed by them.

4.2 Best Sources

As we discussed earlier, we used personal, official, and pay web sites to gather information about our subjects.We found the easiest way to get credible, personal information was to go to our subjects' personal web sites and/or blogs. Personal information gave us current address, telephone, resume (large scale address history), family names, and other information. Personal blogs gave us pointers to other key words or web sites we could visit for more information. The most difficult and time consuming way to get information about our subjects was to look at local media (newspapers, television, yearbooks, etc.) and search for good official pay web sites. Local media web sites typically do not cover the "average joe," thus many of our searches on local media web sites did not come up with anything. The few hits we did get on local media web sites were questionable if it was the same person or if the information was credible since it was from a third party.

Figure 2 shows the size of our fact inventory as it grew over time. The sharp jumps vertically show a quick accumulation of many facts: some of the jumps show the sheer value of some sites. The jump marked A was due to the discovery of Bob's second blog. Our

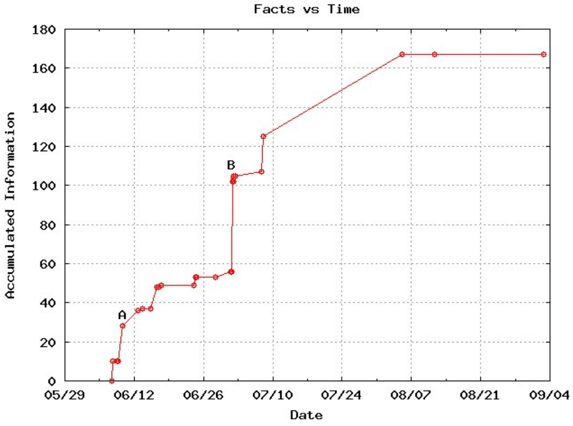

Figure 2: Information Accumulated by Date

inventory jumped at B when we were able to view the Intellius results on both of our subjects.

We believe the subjects' personal websites and Intellius background checks gave us the most credible and largest amout of information. The subjects' personal web sites corroborated information from third party sources (e.g. background checks, professional websites, etc.) We know the Intellius background website is credible because Intellius gets information from public record facilities across the country.

4.3 Pay Site Review

We discussed in previous sections our use of pay sites to get more information about Alice and Bob. We now discuss how much time and money we spent to obtain the 156 facts from pay sites and review the pay sites we used.

Finding a decent pay site was a difficult task. All of the sites promise the most up-to-date information, however none of the sites let us test this claim before giving them anywhere from $10-$100. KnowX (www.knowx.com) ensured users that their databases are updated regularly from official sources, however the web site gave us old addresses for both subjects and did not have any information about two of this paper's authors who have lived in Bloomington, Indiana for at least two years.

We only used pay sites that used secure transactions, but we still had some problems with inconsistent billing by PeopleData (www.peopledata.com). The web site is very easy to use - simply type the person's name in the search field. PeopleData returned with a list of people matching that name and the states they currently live. For $10 users can purchase phone numbers and addresses for all the matches. For $65 users can purchase a public record report about a specific person. For $100 users can purchase the persons public record and criminal report. The reports we purchased were not readily available - they were emailed to us a few days later. The report was similar to other background reports we bought. Unfortunately, we were never emailed or presented a receipt by PeopleData. The credit card charges did not match anything we received. We are still working with PeopleData trying to get charged the correct amount.

The best pay website we found was Intellius (www.intellius.com). Intellius was easy to use and gave us background reports immediately after we paid them. Prices were a little lower than PeopleData and we were emailed a copy of the background report with receipt immediately after our purchase. Intellius had the most comprehensive background report giving us information about aliases, previous addresses, criminal check, neighbors, possible relatives, current house value, and previous residence purchases.

For only $10 we purchased aerial photographs of our subjects' houses.

4.4 Contradicting Information

Not all of the information we found was consistent. Some of the facts presented by even pay-sites was misleading or wrong. It took us some effort in order to discover contradicting and misleading information in the wealth of information we had gathered.

Despite the availability of many addresses, the move-in and move-out dates were often wrong. According to Intellius, Alice has lived in more than ten different places over the last twenty years, yet has only moved out of one location.

Often times these locations are also recorded from a drivers license application, which although it is required by law, some people do not update their license every time they move.

Some people finding sources and web directories often have duplicate listings for the same person. This leads to the question of whether the duplicate is a relative or an alias.

For example, we had originally assumed a name to be Bob's brother. After finding more information, we eventually discovered that it was simply a nickname for the subject. It is sometimes difficult to discern these aliases from relatives.

5 Conclusion

We did not attempt to obtain sensitive documents such as birth certificates and marriage licenses because we would have had to break the law when answering questions about who we were and why we wanted the document. However, many states now offer services to buy copies of such documents online, and it is not clear that appropriate authenticaion mechanisms are in place. Regardless, our Internet Hunt shows that much personal information can be found on the web...legally.

The two best sources were the people themselves and pay sites like Intellius that provide credit and criminal reports for a fee.

In order to address the problems with pay sites releasing information, people need to write their legislative representatives to stiffen laws about the release and selling of such information. Pressure on politicians can increase privacy rights in law. For example, the 1994 Drivers' Privacy Protection act gave consumers an opportunity to forbid the states selling of their driver's license information. This was an opt-out policy that put the burden on consumers. In response to a public outcry in 1999, congress changed the policy from opt-out to opt-in, effectively restricting access to that data.

The law is slow to evolve and it is clear that people could be doing more to protect themselves now. We found blogs, in particular, to reveal an amazing amount of personal information. Both Alice's son and Bob mantained blogs that were close to personal diaries and included names, dates, places, salaries, expenses, hobbies and thoughts. We also found an image of Bob's signature in a scanned in document that he posted to the web, enabling anyone to acurately forge his signature.

In addition to the huge amount of trivia about a person's life which could be used to stalk, aid in identity theft or for other nefarious purposes, our hunt turned up several pieces of information which we consider quite sensitive. Many banking institutions use information such as a person's mother's maiden name, birthdates and home towns as a mechanisms for authentication. We found all three of these with minimal effort. As much of this information was not under the control of our targets, institutions which use such weak authentication mechanisms clearly need to change their authentication procedures.

Until this occurs, however, people need to be educated about the use of this information. The answers to such questions are being used as passwords, making it insecure to use the real information. As a stop-gap protection, people should choose a false and difficult to guess "mother's maiden name". While this puts the burdon on the users to remember the false answers, it may be the only way to protect themselves against such fraud.

Our hunt targeted two specific people with the goal of finding the most information about them. If we did not restrict our targets to specific people, we could have used a different approach and perhaps found more sensitive information. CRM Daily found numerous credit card numbers and associated names and expiration dates by searching for a range of numbers matching credit-card patterns[3]. Other types of sensitive information may be available using similar search tactics.

Much has changed since Rick Gates June, 1993, Internet Hunt. The World Wide Web has evolved considerably. Searching technology is far more advanced; social acceptance and use of the Web has skyrocketed; and most of all the Internet has become an integral part of many people lives. Given these ten years of maturity, the fabric of society has woven deeper into the digital world. Online banking and commerce, widespread use of email, instant messenging, web cams, web boards, blogs, and peer-to-peer file-sharing networks showcase just a fraction of how the Internet has boomed.

With these new technologies, come a new breed of challenges related to security and privacy. Our study illustrated there is a wealth of information available, some with questionable integrity and some with strong ties to personally identifiable information. Given the global momentum to "go digital" and "get connected" this introduces a new level of complexity: cybercrime is now more prevalent than ever. Identity deception/theft, checking account fraud, phishing scams, and computer trespass illustrate this trend.

What will the next ten years hold? Who knows, yet it will undoubtedly be an interesting trip.

6 Acknowledgments

Katie A. Siek is supported in part by a National Physical Science Consortium Fellowship and by a stipend from Sandia National Laboratories/CA. Kay H. Connelly is partially supported by a grant from the Lilly Endowment.

References

[1] Google search engine. Web Site. http://www.google.com.

[2] Yahoo search engine. Web Site. http://www.yahoo.com.

[3] Baker, P., and Baker, B. Google search reveals credit-card numbers. *CRM Daily*. September 16, 2004. (http://story.news.yahoo.com/news?tmpl=story&u=/nf/26967).

[4] Gates, R. June 1993 Internet Hunt. Usenet Post. (alt.internet.services message 8877@ucsbc-sl.ucsb.edu).

[5] Gates, R. June 1993 Internet Hunt Results. Usenet Post. (alt.internet.services message 222sbm$qtecifs2.ucsb.edu).