The Privacy in the Time of the Internet: Secrecy vs Transparency

Murillo Pontual
The University of Texas at San Antonio
mpontual@cs.utsa.edu

Andreas Gampe
The University of Texas at San Antonio
agampe@cs.utsa.edu

Omar Chowdhury
The University of Texas at San Antonio
ochowdhu@cs.utsa.edu

Bazoumana Kone
The University of Texas at San Antonio
bazoumana.kone@utsa.edu

Md. Shamim Ashik
The University of Texas at San Antonio
sashik@cs.utsa.edu

William H. Winsborough
The University of Texas at San Antonio
wwinsborough@acm.org

ABSTRACT

In the current time of the Internet, specifically with the emergence of social networking, people are sharing both sensitive and non-sensitive information among each other without understanding its consequences. Federal regulations exist to mandate how sensitive information (*e.g.*, SSN, health records, *etc.*) of a person can be shared (or, used) by organizations. However, there are no established norms or practices regarding how information that is deemed to be not sensitive may be used or shared. Furthermore, for the sake of transparency, different organizations reveal small amounts of non-sensitive information (*i.e.*, photos, salaries, work hours, size of the houses, *etc.*) about their clients or employees. Although such information seems insignificant, the aggregation of it can be used to create a partial profile of a person which can later be used by malicious parties for robbery, extortion, kidnapping, *etc.* The goal of this work is to create awareness by demonstrating that it is plausible to create such a partial profile of a person just by crawling the Internet. For this, we have developed an open source framework that generates batch crawlers to create partial profiles of individuals. We also show empirical comparisons of the amount of information that can be gathered by using free and also paid websites.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues

General Terms

Security, Experimentation

Keywords

Privacy, Internet, Transparency, Social Networking, Information Sharing

CODASPY'12, February 7–9, 2012, San Antonio, Texas, USA.

1. INTRODUCTION

In the age of Internet, organizations are heavily dependent on computer information systems for using, sharing, and safe-guarding critical information about their clients and employees. Organizations are expected to keep a certain degree of transparency in their functions. Transparency in this case serves several important purposes, such as, creating accountability, building public confidence and trust, and also creating informed stakeholders.

Federal regulations [1, 2] mandate how the organizations can use or share critical private information of individuals. These regulations carry the force of law and violations of them bring in the threat of severe punishment. To maintain a certain degree of transparency, the organizations release some not-so-sensitive information (*e.g.*, name, photo, salary, rent or mortgage, address, office hours, *etc.*) of their clients and employees to the Internet. There is no established norm or custom based on which one can decide how to use or share this information.

This not-so-sensitive information might seem very insignificant with respect to compromising privacy of an individual. However, we argue that this not-so-sensitive information can be aggregated to create a partial profile of an individual, which can then be used by malicious parties to kidnap, extort, rob, *etc.*, the individual (for examples see [3]). This semi-complete profile can also be sold in the black-market.

Social networking [4, 5, 6, 7] has received a lot of attention concerning breaches of privacy of its users [8, 9, 10]. With the emerging trend of social networking, more people are voluntarily sharing not-so-sensitive information (*e.g.*, address, name, email, phone number, pictures, location, *etc.*) with each other without the proper understanding of its implications. Previous work [8, 9, 10, 11, 12] has attempted to create awareness among social networking users by showing the different privacy vulnerabilities that are inherent to these social networking websites, for example by using vacation messages to create lists of vulnerable homes [13] or harvesting email addresses for spamming [12].

As discussed above, previous work is more specifically concentrated on the privacy implications of social networking websites. However, our goal is more general in the sense that we want to create awareness by demonstrating that individuals share a lot more than they need to on the Internet. In this work, we attempt to show the feasibility of aggregating information related to a person by crawling the Internet.

We also try to provide a framework that can be used to generate batch crawlers that can crawl for information in the web.

Contributions. Our **first contribution** is designing and developing an open source framework that creates batch crawlers for different websites in the Internet. The framework learns from a monitored run performed by a user and creates a crawler that imitates the observed behavior on the target website. The tool saves the search results as HTML files which are later processed semi-automatically to gather the data.

After developing the tool, we used it for crawling information about 6000 employees of The University of Texas at San Antonio (UTSA) from the websites White Pages, Intelius, Texas Tribune, *etc.*, [14, 15, 16, 17, 18]. We also calculated probabilities of finding a person's address, age, size of the house, *etc.*, given the name of the person. This is our **second contribution**.

We then bought a subscription for the payed website Net Detective [19] and crawled it for the same information. We compared the information gathered from Net Detective and the free websites. We found out that the information gathered from Net Detective is limited compared to the free websites. A comparison of the availability of the information gathered from the different sources constitutes our **third contribution**.

Road map. The remainder of this paper is organized as follows. Section 2 provides some examples necessary to understand our contributions. Section 3 explains our framework. Section 4 presents the availability of the information collected in free websites. Section 5 presents a comparison of the availability of the information gathered from free and paid sources. Section 6 discusses related work. Section 7 discusses future work and concludes.

2. MOTIVATION

More and more people are sharing public information over the Internet. This trend contrasts with past behaviors, where people did not feel so comfortable about having their information made public, or even being stored in a central database. Several studies (see [3]) before the meteoric rise of the internet in the late nineties indicated that the majority of Americans felt not in control of their information and was against sharing by the government or companies. However, this started to change with the popularization of social networks (*e.g.*, Facebook, Myspace, Orkut, etc.). Several studies [12, 20, 3] have shown that people are unaware of the dangers of making their personal information public over social networks. For instance, according to Garfinkel [3], in a typical social network, 90.8% of the profiles contain photos, 87.8% have a date of birth, 50.8% contain the home address, and 39.9% phone numbers.

Furthermore, not only people are releasing their information, but also governments around the world are making their citizens' personal information public. For example, in Sweden the income tax returns are publicly available over the Internet [21], whereas in Texas, since 1973, by the Texas Public Information Act [14], all the public employees' records are made public.

What kind of harm not-so-sensitive information can bring to a person when it is made publicly available over the Internet will depend on a set of factors, for example, the economical situation of the person, the type of the job the person has, where the person lives, *etc.* For instance, in a developing nation that has many cases of robbery and kidnapping, publishing salaries and addresses is not a good idea.

Given the above motivation we are going to present a few examples that illustrate the dangers of having not-so-sensitive information made public. The following examples were collected from [3, 20, 21].

1. **Direct Marketing** is an advertising technique that utilizes personal information to communicate directly to the customer. Although direct marketing is not a threat *per se*, it can be very aggressive and annoying to the customers.

2. **Stalking, Kidnapping and Theft** are a category of threat where the not-so-sensitive information may be used to harm a person in a real-life situation. An extreme example was the murder of two doctors involved in abortion clinics after their information was made public on a website.

3. **Identity Theft** is the act of impersonating another's identity. There are two ways that one can impersonate another person: one can steal a person's identity in real life, or one can steal a person's virtual identity (*e.g.*, faking a Facebook profile of someone else). With a real-life stolen identity, a perpetrator might for example open new or access existing bank accounts and do monetary damage. A virtual identity can be used for social engineering-based attacks, advertising, slander, cyber-bullying etc.

4. **Phishing, Spam, and Scams** are social engineering techniques that attempt to deceive naive users to release some of their sensitive information (*e.g.*, a credit card number) by making them believe they are communicating with a trustworthy entity. If coupled with identity theft that gains enough information to convince the victim, this is more successful than the generic 419 fraud.

5. **Harm to Reputation** can happen depending on what type of information is made public about a person. A popular example is the screening of applicants by future employers, which today includes Facebook accounts. Here, a person may harm herself. Other examples include vengeful ex-relations sharing personal data, as for example Rate Your Boyfriend [22].

3. DATA ACQUISITION

Most public information is available in the form of online databases. Mainly form-based websites are used to access and search for data. In the age of "Web 2.0", many providers heavily utilize Javascript and AJAX technologies for a smoother user experience without visible HTTP reloads.

To collect the information for our study, we need to access the data stored in the online databases. However, for the number of people we investigate, a manual acquisition is unfeasible. Thus, an automatic way is necessary. Standard crawler software is not applicable to our case: Searching in

such a database does not equate to following links in the form-displaying document.

As a solution, we developed our own crawler framework that is targeted towards form-based sites and supports Javascript and AJAX. The framework is actually a crawler-generator, as we want to crawl many different databases, which each differ in the way forms are designed, search data is entered, and the search submitted.

The framework is implemented as a Java library that models a pipeline for the information retrieval, and uses HtmlUnit [23], which is intended for use with unit testing frameworks. The decision for HtmlUnit was mainly guided by the strong support for Javascript.

We divide the search into five different steps, as shown in Figure 1. The different steps are:

1. Log into the website. Online databases might require an account to access the information. In such cases, we manually register to create an account in that website. After creating the account, the crawler can log into the website to perform any searches. When an account is not required to perform searches, we skip this step.

2. Find the search site. The search form might be on a subpage of the website. For any search, the crawler should navigate to the search page.

3. Perform the search. This includes filling in the form fields and submitting the query.

4. Iterate over results. If the result set of a search is too large, most websites paginate the result. For completeness, the crawler needs to be able to follow pagination links and save all search results to disk.

5. After all searches are completed, the crawler should cleanly log out from the website.

These five steps are connected in a standard order, represented by solid lines in Figure 1. The crawler will loop the pagination step for all sub-pages generated by a query, and will loop steps two through four for all queries.

The dashed lines represent abnormal flow. In case the Internet connection breaks down, the crawler is logged out, or other unusual behavior, the process is repeated from step one.

The pipeline is parameterized by information specific to each crawled website. The Login step requires the address of the login page, as well as the way a login is performed, i.e., which information to put into forms and submit for a login. The Search step usually just needs the address of the search page. The Form Data step needs to know how to split up search information and fill the right form fields. Finally, it needs to submit the search form. The Pagination step needs to recognize pagination links. They are usually formatted in a straight-forward manor. Logging out is usually performed by visiting a certain page.

The most interesting configuration point is step three. On this example the two main problems can be explained: First, the original search parameter (in our case a single name string) needs to be split up for the different search parameters, e.g., first and last name, middle initial, and so on. Second, it is necessary to write that information into the right form fields.

In many cases, the input to the crawler does not have the same structure as the form input on a website. For example, the Texas government employee salary database, as made available, identifies persons by a single string consisting of the first name, the middle initial, and the last name, separated by spaces, whereas those parts are represented by different form fields on most search pages. As we want our crawler framework to be as general as possible, we have to address those structural differences. Our current solution is a lightweight adoption of a technique presented in [24]. The input is tokenized and substrings of tokens can be referenced. For simplicity of the implementation, currently a regular expression is used to correctly split the search strings into tokens under the consideration of a middle initial.

As a result, search information can be stated by information descriptors, which are either string literals, e.g., static login information, or token+substring identifiers. e.g., "#1" for the first token, i.e., the first name, and "#2[1:1]" for the first character of the second token, i.e., the middle initial.

To write the split-up information into the right form fields, we employ a mapping from field identifiers to information descriptors. In most cases, form fields can be uniquely and easily identified by the name tag. In rare cases, further tag information (e.g., the id tag) or even the field type are necessary. Note that not all fields in a form will be mapped, as to not disturb some possible data transfer of the website, like session information.

To ease the use of the framework, i.e., providing the pipeline stages with the right information, we developed a graphical tool that collects the data by monitoring a sample run performed by the user. To this end, we designed an embedded web browser. For a simplified implementation, the instrumentation is done via Javascript event hooks intercepting mouse clicks and URL changes. The form information is collected as plain Javascript helper objects in an array. The clicked form field is also stored as a special element of the array. Finally, the array is serialized to JSON and sent to a small server running in the monitor, which deserializes and stores all interactions.

4. EMPIRICAL EVALUATION

In this section, we summarize our data collection process, sources of data, and also the analysis process of the data collected. We also provide the scope of the data aggregated from different sources.

Table 1: Sources of information (websites)

Data Set	Size	Homonyms Ratio
Texas Tribune	554,790	1:1.132
UTSA Employees	6,316	1:1.0006
White Pages	133,344	1:25.244
Housing	6,596	1:2.956
Net Detective	43,229	1:8.264
Facebook	30,170	1:9.203
Total	750,591	–

4.1 Information Sources

In this study, we have collected more than 750,000 publicly available records containing personal not-so-sensitive information (*e.g.*, name, salary, home address, work department, *etc.*). The first column of table 1 lists the websites

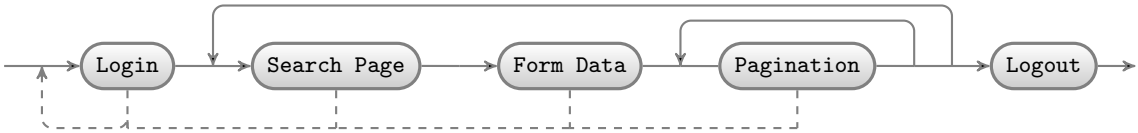

Figure 1: Crawler Pipeline

from which we gathered these information. In the rest of this paper, we call these websites *sources*, *datasets* or *databases* interchangeably. The Texas Tribune [18] is a website from which we collected 667,000 records of government employees from the state of Texas. Typically, these employees are affiliated with public schools, counties, universities, *etc.* A snapshot of an employee record contains the following information: the job title, the public agency where the employee works, the department where the employee is affiliated, the employee gender, the date when the employee was hired, and the employee's annual salary. According to the website, they exclude employees that earn less than $20,000 from their database . Furthermore, when an employee has multiple positions, the salary field contains the sum of all the salaries of such employee, whereas the order fields (*e.g.*, title, agency, and department) contain the information about the highest paid job that such employee hold. The data records are updated periodically according to the Texas Tribune, however, the period is not made available on the website.

The source UTSA employees depicts a subset of the Texas Tribune dataset, containing 6,596 employee records from the University of Texas at San Antonio (UTSA). It contains the same information that was collected from the Texas Tribune website. As presented in the introduction, the goal of this paper is to discover the amount of publicly available information about an individual in the Internet . To this end, in this paper we use the names of the UTSA's employees as inputs to harvest more information in another websites (figure 2 depicts the usage of the names of UTSA's employees as an aggregator to collect more information in other websites).

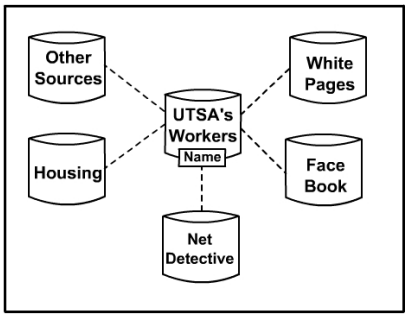

Figure 2: Aggregation method

The White Pages website [16] comprises of the following pieces of information: name, aliases, home address, phone number, relatives, and date of birth (DOB). We have collected 133,344 records of personal not-so-sensitive information by using the names of the UTSA's employees and San Antonio city as inputs to the White Pages engine, .

House Almanac (presented in table 1 as Housing) [17] is a website that contains housing information about some of the cities of Texas (*e.g.*, Austin, Dallas, Houston, and San Antonio). Then, we used San Antonio resident UTSA's employee names to collect housing information for them. We gathered 6,596 records containing the following information: address, co-owner names, subdivision, date of purchase, size, and market price.

Net Detective [19] on the other hand is a paid service that contains the following information about an individual: name, address, aliases, criminal records, phone number, *etc.* We crawled 43,229 records of San Antonio resident UTSA employees from this website. Section 5 contains a more in depth discussion about the information gathered from this website.

We utilized the Facebook engine to crawl for information from the Facebook website. We collected 30,170 records of UTSA employees from Facebook. Each such record contains whether an individual has a Facebook account and the weblink for his profile provided that he has an account. According to [12], a typical Facebook profile contains the following information: name, gender, employers, current city, interests, activities, relationship status, relatives, email, etc. So, our goal when checking whether a given UTSA employee has a Facebook account is a proof of concept that someone can combine the information gathered in free websites (*e.g.*, White Pages, House Almanac, *etc.*) with the information available in Social Networks (*e.g.*, Facebook , Twitter, *etc.*) to build a comprehensive profile about a person.

All the information collected from the sources presented in table 1 comprises of 24 different pieces of information. Together they form a personal profile that details what information is publicly available about a person over the Internet. In addition to that, we have also categorized these pieces of information into 4 main groups, namely, personal information, job information, residence information, and educational information. The categories and respective data items are summarized in Table 2. It is expected that the amount of not-so-sensitive information about an individual will grow to be more complete if we enhance our datasets to contain for information sources.

Table 2: Categories & data items

Personal	Name, aliases, date of birth, age, citizenship, relatives, personal photos, gender
Job	Agency, department, job title, hire date, salary, work phone, work address, email, work history
Residence	Address, subdivision, house size, purchase date, market price, home phone, photos
Education	Degree level, educational history

In table 5 we present an example profile. Note that, not all of the information present in a profile can be acquired from the sources presented in table 1. Information such images, email address, educational information *etc.* were addi-

tionally collected from other sources (*e.g.*, Google, personal websites, Yahoo, *etc.*). Since, our study is limited among the employees of the University of Texas of San Antonio, a good source of information about them were available at the university's official website. We found individual curriculum vitae, class pages, department pages, *etc.*, which were also rich sources of information. Note that, in our example profile we intentionally left out the image field. In section 4.3, we present a more comprehensive study about how this information were gathered.

Finally, we consider homonyms ratio that tells us the number of persons sharing the same name in a dataset. We present the homonyms ratio of the different sources we used, in table 1. This is a rough measure of the difficulty of creating an individual's profile. Recall that we use names of individuals as the matching criterion to aggregate information collected from multiple sources. When a name is too common (*i.e.*, the number of homonyms is too high), one cannot be sure which of the individuals the information belongs to. By inspecting homonyms ratio presented in table 1, we can see that in the Texas Tribune dataset, for each unique employee's name we have 1.132 entries. This a low level of homonyms. The UTSA employees dataset almost does not have homonyms (2 among 6300), and can be used as a good aggregator set. On the other hand, the White Page dataset contains a higher number of homonyms, for each unique name there are 25.244 records. The housing dataset contains a medium number of homonyms, for each unique name, there are approximately 3 records. The Net Detective dataset has approximately 8 records per unique name. Whereas, the Facebook dataset contains 9 records per unique name.

4.2 Analysis

Recall that the goal of this work is to gather as much information as possible about a person from the Internet. With the insights gained above, that on average a name is not unique, we compute the probability of finding certain pieces of information about *any* person with a given name.

To compute the probability to find any information, we count the number of records from a data source that include the data, excluding duplicates as defined by the name. For example, the White pages data source contains age information for some of the records. We count all search results that have the age information, i.e., the age is not given as "unknown", disregarding multiple records for the same name. For instance, the three data records ("John Smith",unknown), ("John Smith",30) and ("John Smith",35) count as one hit for the name "John Smith". We then divide the count by the number of all unique names to compute the percentage of unique names with information, which we consider as an approximation of the probability of finding information given a name.

Table 3 summarizes the probabilities for the different pieces of information associated with a name. We studied the following kinds of information: home address, housing price and size, age, home phone, and Facebook account. However, we should note that the UTSA dataset also implicitly contains information about the gender, hiring date, salary and job description of all persons. Thus the probability of finding this information is trivially 1 for our dataset and omitted for clarity.

The first entry presented in Table 3 is the probability of finding a home address given a name. By inspecting the table, one can see that the probability is very high, 83.6%. However, in addition to the address if one wants to find more details (*i.e.*, house price, house size, house subdivision, date of purchase) about the house associated with a name, then the probability is lower, around 35.3%. The probability for finding the age or date of birth of a specific name is 59.6%. The probability finding the home-phone is 75.0%. Finally, the probability of finding home address, home phone, and age associated with a name is 52.5%. If in addition, one wants to find the age, address, home phone, house price, and house size information associated with a name, then the probability is 26.3%. Recall that, we have a dataset of UTSA employees containing the following information: salary, job title, name, hire date, and gender. Now, if we find any information (*e.g.*, address, age, home phone, *etc.*) then we will also get the information residing in the UTSA dataset.

Finally, given a name, we calculate the probability of finding a Facebook account with that name. Previous studies [20, 12] have shown that people tends to release a lot of personal information over social network websites. They also provided probabilities of finding different fields of the Facebook profile given an email. Once we calculate the probability of a person having a Facebook account, we can reuse the previous work to gather information from Facebook and aggregate the information we collected from the free websites. Including information from Facebook will enable one to create a more detailed (*e.g.*, interests, personality, location, languages, *etc.*) profile of a person.

Our input dataset (UTSA's employees) contained full names including middle initials. A first search was performed using the full information available. If the result is non-empty, that is, Facebook found an account under the full name (in table 3 we call this Facebook full names), we can assume with high probability that this account is for the person searched for. In our study, 25.4% of the dataset had a match with the full name. However, many people do not regularly use their middle initials and omit them in online accounts. Thus, we performed a second search dropping all initials (in table 3 we call this Facebook short names). This potentially results in more matches, but reduces the probability that a search result actually corresponds to the person in question. In such cases, further investigation of the Facebook profile is necessary to increase the confidence in a match. In our study, 39.1% of the dataset had a match with the short name. Overall, the probability of finding a Facebook account given a name (either short or long) is 59.0%.

Table 3: Probabilities of finding personal information from free websites

Properties	Prob.
Address	0.836
Housing price and size	0.353
Address, housing price and size	0.340
Age	0.596
Home phone	0.750
Address, age and home phone	0.525
Address, age, home phone, housing price, size	0.263
Facebook total	0.590
Facebook full names	0.254
Facebook short names	0.391

4.3 Profile Creation

Several interesting properties are hard to collect with automated crawlers, because the search might rely on non-structured websites or complex search results that cannot automatically be analyzed. These include: personal photos, email addresses, relatives, citizenship, work history, house photos, and educational history. To investigate how likely it is to find this information, we manually collected this information for 31 randomly selected UTSA employees, using Google, the UTSA website, Intelius, Yahoo People Search and other sources. The data is shown in Table 4.

The probability of finding personal pictures and email addresses based on a name is very high, 90.3% and 100%, respectively. This results from our selection of UTSA employees, which usually have a work website that contains a picture and address. Equally high is the probability for finding relatives. Many websites like Intelius or Yahoo People Search contain such information. Furthermore, personal documents like dissertations contain family information. The probability of finding a citizenship is around 22.5%. Usually, one can get this information on resumes. The probability of finding where a person had worked before is 87.0%. Multiple sources, most prominently LinkedIn, helps in gathering this information. Online maps are excellent tools for house photos. We used Google Maps, which resulted in a near-perfect probability of finding images of the house or apartment complex. Last, educational history is common to find in our target group (90.3%), since it is contained in resumes.

Table 4: Probability of finding other information

Property	Probability
Personal Photos	0.903
Relatives	1.000
Citizenship	0.225
Work History	0.870
House Photos	1.000
Educational History	0.903
Email address	1.000

4.4 Study Limitations

All the probabilities computed in this section considered that given a name one can find some information about the name. However, we cannot guarantee that when we aggregate data from different sources, the aggregation will result in a profile that is related to exactly the same person we have queried for. In fact, since each dataset contains more than one record per unique name, the probability that we get gives us only the chance of finding a match between the name and desirable information. Although, this is not ideal, one can try to find clues in other websites to increase the confidence in a match. For instance, the housing dataset contains information about the owners of a house, in this case, this information usually contains the names of a couple. As the White Pages dataset also contains names of relatives of an individual, one can use the spouse name to narrow down the search result and increase the confidence of a match.

In addition to this, we use as our aggregator dataset the names of the UTSA's employees. By doing this, we also got the job title, gender, salary, and hire date. However, we understand that not all state governments and organizations release this type of information over the Internet. Two opposite examples are Sweden and Brazil. In Sweden, personal tax return information about every citizen is made available online for public use. This of course includes job titles and salaries, and with a history search also hiring dates. Brazil, on the other hand, while having a very efficient tax system, does not release any information publicly. To bridge this gap in release policies, we note that usually third-party information providers are available. For example, the Glassdoor [25] website provides average salaries for a diverse range of companies in the US. This information can be used to approximate missing data points. In that sense, we feel our approach can be successfully extended to different types of users, even though they do not work in a country, state, or company that publishes all of their information publicly available online.

5. COMPARISON WITH PAID ALTERNATIVE

The Net Detective website is a paid service that claims to offer more than 1.1 billion records about people around the world, and 231,461,546 records for US residents. However, the amount of records that could be collected from it when using the name of UTSA's employees is around 1/4 of the size of what we could collect from the White Pages records (see table 1). During the course of our study, they charged $29.00 for unlimited access for 3 years. Their website advertises that one can find the following pieces of information about a person: people searches, criminal records, sex offender list, arrest records, phone records, address records, social security records, public records, birth, marriage, divorce, and death records. However, what we found was a different story. The $29.00 subscription only allowed us to check for home address, date of birth, and home phone. They charged an additional $10.00 non-refundable processing fee for 3 days of accessing other additional information. We did not pay this fee considering it to be a scam as they did not live upto their initial promise of providing the information that comes with the $29.00 subscription fee.

5.1 Analysis

We follow the same methodology described in section 4 to compute the probability of getting information about individuals from the Net Detective website. Table 6 summarizes these probabilities. Given a name, the chance of getting a full home address containing zip code, state, city, and street is 82.8% which is very similar to the probability for White Pages (83.6%). The probability of finding a date of birth in Net Detective is 35.7%, whereas it is 59.6% for White Pages. Similarly, the probability of finding a home phone is 68.7%, which is less than that of the White Pages (75.0%). Finally, the probability of finding these 3 information together is 27.7%, which is 52.2% for the White Pages. To conclude, we felt that this site is not a good alternative to collect public information about people. In an approximation, the information we gather from the Net Detective service was a subset of the information we collected from the free service of White Pages.

6. RELATED WORK

David Brin is one of the first researchers to study the relation of privacy and its relationship with modern soci-

Table 5: Profile Example

Personal Information			
Name	Alice Eve	**Alias**	Ali Eve
DOB	10/09/2007	**Age**	42
Citizenship	American	**Relatives**	Bob Eve (Husband)
Gender	Female		
Job Information			
Company	UTSA	**Job Tittle**	Secretary
Hire Date	11/12/2005	**Salary**	$45,000.00
Work Phone	(210) 444-4444	**Email**	asmith@utsa.edu
Work History	UT Tyler, secretary 2004 - 2005		
Residence Information			
Address	111 Flower Street, 78221	**Subdivision**	Flower
Size	1200 Sq feet	**Date Purchase**	06/02/2006
Current Price	$130,000.00	**Phone**	(210) 222-2222
Educational Information			
Degree	B.A. in Geography	**History**	Bachelor in Geography, UT Tyler - 2000-2003

Table 6: Probabilities of finding personal information in the Net Detective website

Property	Probability
Address	0.828
Age	0.357
Home phone	0.687
Address, age and home phone	0.277

eties [21]. In his opinion, the only viable option is to make all personal sensitive information public, in his words "A transparent society". In similar lines, Garfinkels [3] shows how our society is already becoming a transparent society. These two books are more philosophical, whereas our work is more practical. We collect real public data to see how much information can one gather about a person via the Internet. In other words - how transparent is our society.

Many researchers have been studying the impact of social networks on the reputation and privacy of its users. In particular, Jin *et al.* [26] presents techniques for detecting whether a user profile in a social network is cloned, whereas Solove [20] shows more ethical and philosophical issues about the topic. In contrast, the scope of our work is broader and it discusses how someone can learn about others over the Internet (not restricted to social networks). In that sense our research is complimentary with [20, 26].

Harvesting information on databases is a well known topic in the literature, and it has raised a lot of discussions about ethical questions [27, 28, 29, 30]. However, few researchers have presented techniques for harvesting personal information over Social Networks. In particular, Polakis *et al.* [12] present a comprehensive study how someone using an email address as input can crawl personal information on Facebook, and other social networking websites. On the other hand, Krishnamurthy *et al.* [31] point several techniques that can be used to leak sensitive personal information from social networks. The work in [12] is the one that is more similar to ours. Nevertheless, where that work is concentrated on using Facebook profiles given an email to harvest personal information, we use a series of free websites to gather other types of information that were not considered in [12]. In that sense, our work is complimentary also to their work.

In a previous section, we have described how personal information gathered from the Internet can be used by malicious parties for different purposes. One of the threats that has received more attention from the researchers is the Spam attack [32, 33]. Compared to our work, both researchers focus more in defining how spammers act, and how they collect the information that will be used to send Spam emails, whereas the scope of our work is more general.

7. CONCLUSION AND FUTURE WORK

In this work, we observe that with the growing popularity of social networking websites, people are sharing not-so-sensitive information (*e.g.*, location, images, employer name, phone number, email address, home address, personal websites, *etc.*) among each other without understanding the consequences. Furthermore, for the sake of transparency organizations also release some information of their customers or employees that is deemed to be not-so-sensitive (*e.g.*, name, salary, home address, work department, *etc.*). However, individuals are not fully aware of all the information that is available about them on the Internet and even if they are aware, consider this information to be insignificant. Although the individual pieces of information seem harmless, the aggregation of this information can be used against the individual for kidnapping, robbery, extortion, spamming, phishing, direct marketing, *etc.* We show that it is possible and feasible to aggregate this information about a large number of people just by crawling the Internet. To this end, we have developed a framework that creates batch crawlers that imitate and automate a user's interaction with a website to crawl information about many individuals. We used the framework to crawl a multitude of *free* and *public* data sources like social networking sites, and also included a paid alternative. The crawled data was then used to compute the probabilities to find certain kinds of not-so-sensitive information that together can potentially become dangerous. Included in the analysis is also a surprising comparison of the free and for-pay alternatives: it turns out that the available free information is already stronger than the paid-for version.

The result of this study is a number of significant probabilities of finding sets of certain kinds of information. For example, the probability to find all of the gender, salary, address and housing price of a person is greater than one in three. While we do not intend to give advice or suggest social norms, we believe those numbers should be taken seriously. Internet users should be aware of the potential side effects and amount of publicly released data, because only

informed netizens can make informed decisions about their behaviour, the sharing of information, and the associated risks on the Internet. From our experiments, we realized that without significant resources (except for time) and with the right tools, gathering not-so-sensitive information about many individuals is not that challenging.

Future work. As our future work, we intend to perform a comprehensive user study that will expose individuals to our findings. Namely, we are interested in the reactions of participants to the amount and detail of information that is available about them on the Internet, and whether there should be federal regulations that would mandate the amount of not-so-sensitive information that can be revealed by organizations for the sake of transparency. A small-scale pilot study seems to confirm our suspicions about the average Internet user: it revealed that the participants were usually astonished by the amount of personal information that was available on the Internet about them. Furthermore, we want to extend our analysis for other countries (*e.g.*, Brazil, Sweden, *etc.*) and compare their views with respect to privacy to the views taken in US.

8. REFERENCES

[1] "Health resources and services administration," 1996, health Insurance Portability and Accountability Act, Public Law 104-191.

[2] "Federal trade commission, How to comply with the children's online privacy protection rule," 1999, public Law.

[3] S. Garfinkel, *Database nation: the death of privacy in the 21st century*. Sebastopol, CA, USA: O'Reilly & Associates, Inc., 2001.

[4] "Facebook," http://www.facebook.com.

[5] "Myspace," http://www.myspace.com.

[6] "Orkut," http://www.orkut.com.

[7] "Hi5," http://www.hi5.com.

[8] K. Thomas, C. Grier, and D. M. Nicol, "unfriendly: multi-party privacy risks in social networks," in *Proceedings of the 10th international conference on Privacy enhancing technologies*. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 236–252.

[9] D. Freni, C. Ruiz Vicente, S. Mascetti, C. Bettini, and C. S. Jensen, "Preserving location and absence privacy in geo-social networks," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. New York, NY, USA: ACM, 2010, pp. 309–318.

[10] A. L. Young and A. Quan-Haase, "Information revelation and internet privacy concerns on social network sites: a case study of facebook," in *Proceedings of the fourth international conference on Communities and technologies*. New York, NY, USA: ACM, 2009, pp. 265–274.

[11] A. Ho, A. Maiga, and E. Aimeur, "Privacy protection issues in social networking sites," in *Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference on*, may 2009, pp. 271 –278.

[12] I. Polakis, G. Kontaxis, S. Antonatos, E. Gessiou, T. Petsas, and E. P. Markatos, "Using social networks to harvest email addresses," in *Proceedings of the 9th annual ACM workshop on Privacy in the electronic society*, ser. WPES '10. New York, NY, USA: ACM, 2010, pp. 11–20.

[13] "Please Rob Me," http://pleaserobme.com/.

[14] "Texas Public Information Act-1973," https://www.oag.state.tx.us/open/.

[15] "Intelius," http://www.intelius.com/.

[16] "White Pages," http://www.whitepages.com/.

[17] "House Almanac," http://www.housealmanac.com/.

[18] "Texas Tribune," http://www.texastribune.org/.

[19] "Net Detective," https://www.netdetective.com/.

[20] D. J. Solove, *The Future of Reputation: Gossip, Rumor, and Privacy on the Internet*. New Haven, CT, USA: Yale University Press, 2007.

[21] D. Brin, *The Transparent Society: Will Technology Force Us to Choose Between Privacy and Freedom?* Perseus Books Group, Jun. 1999.

[22] "Rate Your Boy Friend," http://www.rateyour-boyfriend.com/.

[23] "Htmlunit," http://htmlunit.sourceforge.net/.

[24] S. Gulwani, "Automating string processing in spreadsheets using input-output examples," in *Proceedings of the 38th annual ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, ser. POPL '11. New York, NY, USA: ACM, 2011, pp. 317–330.

[25] "Glassdoor," http://www.glassdoor.com/.

[26] L. Jin, H. Takabi, and J. B. Joshi, "Towards active detection of identity clone attacks on online social networks," in *Proceedings of the first ACM conference on Data and application security and privacy*, ser. CODASPY '11. New York, NY, USA: ACM, 2011, pp. 27–38.

[27] H. T. Tavani, "Informational privacy, data mining, and the internet," *Ethics and Inf. Technol.*, vol. 1, pp. 137–145, January 1998.

[28] L. Van Wel and L. Royakkers, "Ethical issues in web data mining," *Ethics and Inf. Technol.*, vol. 6, pp. 129–140, June 2004.

[29] M. S. Olivier, "Database privacy: balancing confidentiality, integrity and availability," *SIGKDD Explor. Newsl.*, vol. 4, pp. 20–27, December 2002.

[30] A. Narayanan and V. Shmatikov, "How to break anonymity of the netflix prize dataset," *CoRR*, vol. abs/cs/0610105, 2006, informal publication.

[31] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identifiable information via online social networks," in *Proceedings of the 2nd ACM workshop on Online social networks*, ser. WOSN '09. ACM, 2009, pp. 7–12.

[32] C. Kreibich, C. Kanich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage, "On the spam campaign trail," in *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*. Berkeley, CA, USA: USENIX Association, 2008, pp. 1:1–1:9.

[33] M. B. Prince, B. M. Dahl, L. Holloway, A. M. Keller, and E. Langheinrich, "Understanding how spammers steal your e-mail address: An analysis of the first six months of data from project honey pot," in *CEAS*, 2005.