

Profiling Game

Claudia Díaz

Abstract—“How much information on someone can you find in the Internet?” This question motivated a profiling game in which 15 PhD students participated. This report describes how the game was conducted, the tools that were used to obtain information on the profiled subjects, the information that was found, the methodologies followed by the participants in order to build profiles and discussion and conclusions on the game.

I. INTRODUCTION: THE GAME

This document describes a profiling competition that took place in January 2005 as part of the program of the First FIDIS PhD Event. The competition was organized as follows: fifteen PhD students, divided in three groups of five people, had to collect as much information as possible about three target subjects, using for this purpose information publicly available on the Internet. The participants had a few hours to complete the task.

The purpose of the game was to use this simple experiment to get an idea on the amount of data available on the Internet, as well as to explore approaches to collect and combine these data. The participants were given freedom to decide how to proceed with the collection and combination of data, as far as only publicly available data was collected. In order to be able to verify the correctness of the collected data, the target subjects were among the participants (one per group), and each group had to collect data on the two target subjects who belonged to the other groups. As a disclaimer, it is worth nothing that the specificity of the profiled subjects (German PhD students with active Internet lives), does not (yet) make these results generalizable to the information available on an average citizen.

At the end of the game, the groups had to present the profiling information of the target subjects they had studied. Each piece of data had to be linked to the web address or resource where it had been obtained. Groups obtained points for the data that was found exclusively by them. When a group had mistakenly added to the profile information that was not related to the subject (e.g., data belonging to someone else with the same or a similar name), points were lost.

In the following sections, we describe the tools used to collect the information, the type of information that was found, the strategies used to build the profile, the main issues addressed by the discussion that took place after the game and the conclusions we have extracted from this experiment.

II. TOOLS TO COLLECT DATA

The most heavily used tools for the collection of data were search engines. They were used to get the basic information, that was later complemented with other search tools. The participants reported having used the following tools:

- Google was extensively used. The keywords searched were initially name and surname. Later on, as more information on the subject became known, other keywords such as town or email address were used in order to refine the search or obtain additional information.
- Other search engines, such as Altavista provided complementary information. Search engines for a specific type of content, such as images.google.com or groups.google.com also provided complementary data (like pictures) on the profiled subjects. Google News was used to find newsgroup posts from student times.
- National and local phone books. As the nationality of the three profiled subjects was German, the German phone book proved to be useful to provide the home address, the phone number and sometimes even the profession of the subjects. Searches for people in the same town with the same surname led to the discovery of relatives.
- Home pages proved to be a rich and reliable source of information, they typically provide a CV that describes many of the activities developed by the subject and the places and institutions in which he or she has worked. Moreover, it usually contains some indications of the preferences and interests of the subjects, as well as a publication list. Once the subjects were linked to other people, the home pages of co-authors, friends, colleagues and relatives provided additional information.
- Archived sources such as google's cache or www.archive.org provided information that was no longer available in the original site, either because it had been removed or because the website was no longer active.
- Various messenger accounts were used to obtain information; for example ICQ, Skype.com, MS Messenger and Jabber. Orkut and Open BC Membership provided information on the social network and the interests of the subject.
- OpenBusinessClub provided information about membership in organizations.
- Key servers provided the PGP key and information on the social network of the subjects (by looking at the people who had cross-signed their key with the subject).
- School pages, once the schools the subject has attended become known (typically mentioned in the CV).
- Domain's owners search (www.denic.de) and search for home pages in web domains that contain the name or surname of the subject. This provides additional home pages the subject may have or maintain.
- CiteSeer, DBLP and other publication search engines provided the publications of the subject and links to co-authors and related researchers.
- Newspapers, mainly local newspapers from the home

town of the subject provided information on their public activities. Some groups also searched for events that were reported on the birth date of the subject.

- Newsgroups on topics of interest for the subjects provided information on the technical activities and knowledge of the subjects, as well as relationships.
- Well known stores, such as Amazon.com, were looked into in order to find information on the books, films, or other articles the subject buys on the Internet.

Some participants proposed a list of additional tools that could be used to further enhance the profiling, but which were not used due to the limitation of time for the exercise. Additional data gathering tools include:

- Buying the financial record from Schufa.
- Buying the consumer profile from Informa.
- Run a password guessing tool for web form logins on Amazon and eBay with the subject's e-mail-addresses to harvest the business histories these sites keep.
- Inquire the subject's highschool, university and employer to ask about certificates, graduation etc.
- Check the many find-your-classmates web pages.
- Write him/her an email to get one back to see the IP-address and the Mail-Client.
- Try to sniff the communication of the subject in order to monitor his or her Internet activity.

III. RESULTS OF THE SEARCH

The first goal of the participants was to find personal homepages. These were easily found by searching the names in Google, as they appeared on top of a list of a few hundreds of hits. The home pages were used to collect a basic set of information such as:

- Work place, position, professional activities, work location
- Contact information: town, address, phone, fax, email address
- CV, which typically contains educational and professional history of the subject (name of previous companies, schools, universities, etc.)
- Publication list, which provides information on the research interests and co-authors
- Picture (a picture of the subject is often available)

From this basic information, and using the tools listed in the previous section, the participants were able to obtain for one or more of the profiled individuals additional information such as:

- Personal phone, address, alternative email addresses
- Languages spoken
- Political affiliation, and participation in political activities or discussions
- Religious affiliation, and participation in religious groups
- Participation in public awareness campaigns in the local town
- Multiple alternative email addresses that served as pseudonyms in certain domains. These email addresses could have been further investigated in order to find additional information related to the subject under the pseudonym

- Pictures of the subject and taken by the subject
- Military membership and grade
- Date and place of birth
- Name, address and phone numbers of relatives
- Sports, hobbies and miscellaneous interests (photography, philosophy, etc.)
- Social network: names of the co-authors, colleagues, friends (including the name and picture of an *Internet fan*) and relatives of the subjects were easily found
- Places the subject has been to and dates of the stays
- Personal diaries full of personal details (in some cases travel diaries) revealed daily patterns of activity
- Opinions posted in a variety of Newsgroups
- An image of the (physical) signature

The amount of misleading data (which was unrelated to the subject) found was minimal. However, it is expected that searches on subjects with more common names than the profiled ones in this particular game would generate larger amounts of misleading information (or would require more effort in order to separate the information that refers to the subject of interest and that belonging to other, unrelated subjects).

IV. BUILDING A PROFILE

The limited duration of the game, and the large amounts of data to be gathered, left little time for sophisticated profiling of the subjects. Most groups presented the collected data, classified as belonging to distinct areas of identity. This organization of the data led to some preliminary conclusions on meta-data, which is not directly available but inferred by the combination of several sources of data.

The subjects were naturally implementing identity management mechanisms, as the information they provided varied according to the topic and environment of the hosting web site. For example, one of the subjects had provided an informal picture for his student website, and a formal one for his profile in a business website. However, all these data could be linked together with simple searches in the Internet (as had been made obvious by the first part of the exercise). The distinct profiles that could be built on each of the profiled subjects were:

- The **educational** and **professional** profiles were easily available for any of the subjects. Detailed and complete information was provided in the personal homepages. This is consistent with the idea that it is in the interest of the subject to have a publicly available professional identity, because it is useful to improve his or her professional career. Even more, we could say that a public professional profile is today required in the research field in order to conduct daily working activities.
- Several aspects of the **personal** profile were also easily available. In particular, there was extensive information related to hobbies, sports and other interests (some somehow related to the professional activity, like participation in Newsgroups on topics related to their work; and some totally unrelated, like photography or philosophy). Much of this information was made available through the homepages (or pages linked to it), and other was easily linkable to the subject, as it usually contained his

or her name and other verifiable information. From this observation, we could say that users often like to make public these details, possibly as means to be reachable for people who share similar interests. More sensitive information, like sexual orientation, religious beliefs or attitudes towards drugs could also be extracted from the available data.

- Unfortunately, there was no psychologist among the participants. However, a draft **psychological, behavioural and ideological** profile could be built, both based on the opinions (and the websites in which opinions were posted), language use, pictures (face, expression, clothing, background) topics discussed, and personal experiences described in Newsgroups, posts and online diaries (which proved to be a very rich source of information for building a psychological profile). Some of the groups selected some personal details for their expositions in order to give a feeling on the degree of detail of some data that can be found and linked to a person.
- Most of the groups presented names of people who appeared to be related to the studied subjects. Participants highlighted that, if more time was available, it would be easy to build the **social network** of the subject. Names and addresses of relatives were found in the phonebooks; names of colleagues and co-authors were available in the homepages; some groups tried to google the address or phone number to look for other people living with the subject; Membership accounts provided social contacts of the subjects; descriptions of trips or events often included names of other people; and key servers provided the trusted relationships with other PGP users.
- One of the aspects that proved harder to find with the available tools was the **consumer** profile. Some groups could withdraw conclusions on items of interest for the subject from the description of personal interests or experiences, as well as from the available pictures. However, a more detailed search may reveal the consuming patterns of users who have an account in eBay, Amazon, or other e-commerce websites. Moreover, it is worth noting that this information is not available to the Internet user, but it is to the companies with which the subject has made transactions (and to all those other companies that buy the data share information with the original company). In this sense, the data is in fact available to those who are interested in exploiting it in order to build -valuable-consumer profiles.
- No **financial** information was available through search engines. More sophisticated techniques (possibly involving social engineering or hacking activities) may be required in order to access this sort of information. However, the financial status of the subjects could be roughly estimated from the data. Similarly, no **health** related information was found. Nevertheless, assumptions could be made on the general health condition and the lifestyle of the subject by examining the available information (for example, someone involved in hard sport activities could be assumed as being in a healthy condition).

The combination of information related to a subject, belonging to different contexts, opened the possibility of sophisticated profiling. When many different identities of the profiled individuals were put together (combination of professional, personal, and leisure activities), the result was a rather detailed and complete picture of the subject. Even though this was only outlined by the participating groups, the large amounts of data gathered (which could not be thoroughly analyzed) suggested that building a sophisticated profile, which included most important aspects of a person's life, was possible.

V. DISCUSSION

A short discussion followed the presentations of the profiles. One of the issues that was raised was the very question of the meaning of profiling. While some argued that the fact of putting together facts relative to a subject which comes from different sources is effectively building a profile, others argued that the gathering of information should be understood as *data mining* and that *profiling* consisted in inferring meta-data from the raw data (i.e., interpreting the available data in order to classify the subject according to abstract models or categories).

The freedom given in the definition of the game "*build a profile as complete as possible*", led to different interpretations. While some groups looked for as many details as possible, others focussed on more general information that could give a "big picture" of the subject. Similarly, not every group used the same searching tools. Some focussed on the social network, some others on the professional and educational profiles, some on the social activities, and some on their writings. This led to another relevant question: What is relevant to build a profile? In order to answer this question, we need to know what is the purpose of building a profile. Clearly, a company wanting to sell a product is interested in profiles which provide the information of how likely is a customer of buying a certain product; while a potential employer will be interested in the educational and professional profile.

Given that the profiled individuals were among the participants, we could ask the question of whether they were really aware of the amount of information related to them that was publicly available in the Internet. They acknowledged that, although much of the information had been made public by themselves, they were not aware of the existence (or linkability to their public identity) of certain data (in some cases, such as contact details of relatives, the data could seem scary).

What are the potential threats a person could face due to the exposure of personal information? It is clear that an adversary wanting to harm the reputation or even the physical life of a subject, could use the Internet as a source of valuable information. In another level, companies can use this information in order to sell their products or services to the subject, organisations seeking to extend their membership (e.g., religious sects) may use this information in order to target individuals who are more likely to be converted, and public powers may also profile their citizens in order to exercise more control on the populations (e.g., anti-system individuals could be easily monitored and controlled).

Why do people provide information on themselves? The main advantage and motivation for making personal data pub-

licly available is that having a persistent identity in the Internet is a necessary condition to develop professional, personal, commercial and leisure activities. Offering a public profile enhances accessibility for other people who share interests. Keeping a public profile is also a tool to build reputation (e.g., providing the CV and the list of publications). A public identity is, in short, necessary to develop enhanced professional and personal activities, as well as a tool to extend the social network beyond traditional territorial borders.

VI. CONCLUSIONS

Here we summarize some of the most important conclusions extracted from the experiment:

- For individuals who develop activities in the Internet, large amounts of information linkable to their identity are available. This information includes professional, educational, personal and psychological data, as well as enough information to build (at least) a partial social network.
- Sensitive data such as financial, consumer and health information, religious and political beliefs or sexual orientation are harder to find, but it either can be found with more sophisticated search tools or inferred from the available information.
- Internet users have a clear interest in making public their educational and professional profile in order to widen their professional opportunities. Regarding the personal profile, some aspects are also made public, possibly to share information or establish communication with people who share similar interests.
- There are certain aspects of the personal profile for which the subjects do not have an interest in making easily linkable to their professional profile. Yet, this information is easily linkable with simple search tools (due to the fact that it shares an identifier, e.g., the name, with the professional profile).
- Profiled subjects found most disturbing the public availability of data they had not introduced in the Internet and which referred to their *offline* lives (for example, the home phone number or address), mainly that available in public indexes as phonebooks. Also, subjects seemed to be surprised by the possibilities of constructing a social network, which ranged from names and locations of relatives to an extensive network of social and professional relationships which could hardly be denied.
- Subjects were performing some identity management when interacting with different parties and in distinct contexts. Yet, many of these partial identities could be traced and linked to a natural individual for which extensive information on his work, life, interests and location was available.
- People (including technically educated people) do not seem to be fully aware on the information on themselves that is publicly available in the Internet and easy to find and link. When they realize the degree of public exposure to which they are subject, there is often a feeling of vulnerability.
- The exposure of personal data may constitute a security problem. On the other hand, it is used as a tool to build reputation, enhance the professional and personal network, and share information with people with similar interests.
- Once information has entered the Internet, it cannot be removed. Web pages which are no longer available can still be retrieved from archives and caches. Therefore, once the data has been introduced in the Internet, it is no longer under the control of its owner.
- Human intelligence was behind the search strategies and profiling methods. Although automated searches could be helpful in the collection of data, complex decisions need to be taken in order to discriminate relevant from irrelevant (and even misleading) information. This high cost in time and human resources for building just one profile may discourage massive sophisticated profiling.
- The intervention of human intelligence may explain the minimal amount of misleading information that was presented as part of the profiles.
- Social networks, on the other hand, are easier to build in an automatic way. A program could easily search for links in webpages and appearances of identifiers such as names, email addresses or locations in order to construct a social network.
- The results provided by the search engines were much better for language dependent content. Regarding the language issue, it also seems that people (at least Germans), even if they have good knowledge of English, tend to write in their mother tongue.
- Finally, there is clearly a need for identity management tools that allow for pseudonymous, unlinkable management of information that belongs to separate contexts in order to empower the user in the management of his or her data and identity.