Using Social Networks to Harvest Email Addresses

Iasonas Polakis
polakis@ics.forth.gr

Georgios Kontaxis
kondax@ics.forth.gr

Spiros Antonatos
antonat@ @ics.forth.gr

Eleni Gessiou
gessiou@ics.forth.gr

Thanasis Petsas
petsas@ics.forth.gr

Evangelos P. Markatos
markatos@ics.forth.gr

Institute of Computer Science, Foundation for Research and Technology Hellas
Heraklion, Greece

ABSTRACT

Social networking is one of the most popular Internet activities with millions of members from around the world. However, users are unaware of the privacy risks involved. Even if they protect their private information, their name is enough to be used for malicious purposes. In this paper we demonstrate and evaluate how names extracted from social networks can be used to harvest email addresses as a first step for personalized phishing campaigns. Our *blind harvesting* technique uses names collected from the Facebook and Twitter networks as query terms for the Google search engine, and was able to harvest almost 9 million unique email addresses. We compare our technique with other harvesting methodologies, such as crawling the World Wide Web and dictionary attacks, and show that our approach is more scalable and efficient than the other techniques. We also present three *targeted harvesting* techniques that aim to collect email addresses coupled with personal information for the creation of personalized phishing emails. By using information available in Twitter to narrow down the search space and, by utilizing the Facebook email search functionality, we are able to successfully map 43.4% of the user profiles to their actual email address. Furthermore, we harvest profiles from Google Buzz, 40% of whom provide a direct mapping to valid Gmail addresses.

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General; K.4.1 [**Computing Milieux**]: Computers and Society

General Terms

Security, Measurement

Keywords

Social Networks, Email Harvesting, User Profiling, Spam Email

WPES'10, October 4, 2010, Chicago, Illinois, USA.

1. INTRODUCTION

Online social networks (OSNs) have attracted the interest of millions of users. Facebook has more than 400 million users [3] while Twitter has more than 40 million users (as of July 2009) that exchange over 50 million tweets per day [16]. Users are able to interact with each other, chat, share thoughts and links, play games and conduct several other activities. The popularity of social networking has also attracted the interest of the research community that tries to understand their structure and user interconnection [23, 28] as well as interactions among users [29].

As users tend to share personal information and activities, privacy leakage is one of the biggest problems of social networking. Personal information is not limited to a name, birth date, religion and marital status. Participation in events, friend lists, groups and organizations the user belongs to, preferences in music and food also reveal information about the life of the user. Articles revealing stories of employees losing their job or not getting hired due to information contained in their Facebook profile have gained wide attention [2, 10]. Even though in some networks users can fine tune their privacy settings (they have the option to share their personal information only with their friends, up to second degree friends, their network, everybody or nobody), information leakage still remains an important problem as many users do not always understand the implications of revealing personal information online [12].

Social networks can become a valuable resource for attackers. In earlier work it has been demonstrated that attackers can impersonate users in order to steal private information [20]. Privacy leakage attacks [31] can be used in many ways, such as revealing sensitive information for "high value" targets. One of the most sophisticated attacks based on harvested private information is personalized phishing. In traditional phishing schemes, emails contain generic terms, such as "Dear user", "Dear customer", "Hello subscriber" etc., which may be considered suspicious by many of the targets. Personalized phishing follows a different approach. The emails are crafted in a way so as to look like they originate from a friend or a relative of the potential victim. This type of email is far more convincing than the classic 419s scams [27] as it directly addresses the recipient and appears to be sent from someone the victim knows.

In this paper we demonstrate that social networks are an enormous and ever expanding pool of information that can be used as a stepping stone for personalized phishing campaigns. We demonstrate that even by retrieving the most

basic information, i.e. the name of the user, we are able to harvest millions of email addresses. We present two different approaches to harvesting; *blind harvesting* that aims to gather as many email addresses as possible, querying for names retrieved from OSNs in the Google search engine, and *targeted harvesting* that aims to gather email addresses and correlate them to personal information publicly available on social networking sites.

Using the *blind harvesting* methodology we were able to harvest, on average, 45 emails per name for the Facebook names and 25 emails per name for the Twitter nicknames. Our results show that this approach can harvest more addresses than traditional harvesting techniques in a highly automated, scalable way that requires little runtime and network overhead.

We present three *targeted harvesting* methodologies. The first uses the email-based search capability of Facebook. We collect names from highly populated Facebook fan pages and use the blind harvesting technique to search for email addresses. We then use the harvested email addresses in the Facebook search utility. If one or more profiles are returned, we check whether any of them have a matching name to the one collected from Facebook and map them to the email address in question. Such confirmation allows the use of personal information available in that profile to craft a personalized phishing email. This correlation technique can successfully link 11.5% of the harvested names with their actual email address. In order to improve the efficiency of the first technique, our second technique uses information from the Twitter network. By collecting <nickname,name> pairs from Twitter, we harvest emails with a prefix that is an exact match to the nickname and then search for them in the Facebook network. This technique can successfully correlate a user's profile with his email address for 43.4% of the profiles returned as part of this Facebook lookup. Our last technique relies on searching Google Buzz, using the names of users collected from other social networks, to discover profiles and additionally crawl through their follower relations. Our experiments showed that 40.5% of the Buzz profiles we collected revealed the user's account name, which is also the user's Google mail account. Thus, by using our technique, one can harvest the actual email address of the targeted user and all the personal information that is revealed in their Google profile and Buzz posts.

The rest of this paper is organized as follows. Section 2 outlines the related work, while in Section 3 we present challenges that are inherent to social networks. In Section 4 we present traditional harvesting methodologies, and in Section 5 we describe in detail our harvesting methodologies that use social networks. Our measurements are analyzed in Section 6 and discuss defensive countermeasures in Section 7. We refer to future work in Section 8 and conclude in Section 9.

2. RELATED WORK

To understand the ways that spammers obtain target email addresses, Shue et al. in [26] post a number of email addresses on popular websites and monitor the inflow spam that these addresses receive. They also conduct a study of the current web crawlers that spammers are equipped with. The results of this work led the authors to three major conclusions; first, email addresses are discovered quickly on the Internet by spammers. Second, spamming crawlers can be tracked and, finally, most spammers use multiple email-harvesting techniques on a plethora of sources including web pages, blogs, social networking sites, mailing lists etc. According to the authors, even a single exposure of an email address can result in instant and high-volume spam.

Prince et al. [25] present the results of Project Honeypot [13], which aims to reveal the primary way by which spammers collect new email addresses. As the authors state, harvesting is the basic methodology used by spammers to obtain new email addresses. They divide harvesters into two classes according to their turnaround times from the moment an email address is harvested until the first message is sent. *Hucksters*, are characterized by a slow turnaround time, while *fraudsters* send the first message almost instantly after the address is harvested.

Kreibich et al. [22] try to fully measure the orchestration of spam campaigns by hooking into botnet command-and-control (C&C) protocols. Among others, the authors make an analysis of hundreds of millions of harvest reports (lists of target email addresses that spam bots harvest) that were collected through their proxies. They observe that the most frequently harvested domains correspond to major email services such as *hotmail.com, yahoo.com, aol.com*. Furthermore, almost 10% of all harvested email addresses do not correspond to valid top level domains.

Krishnamurthy and Wills [24] describe how third-party servers can exploit the personal identifiable information (PII) leakage of social networks so as to link it with user actions inside these networks or even elsewhere on the Internet. The authors demonstrate that most users can have their PII linked with tracking cookies. They state that this is a corollary of users' ignorance about the importance of strong privacy settings in a social network.

Bilge et al. [20] show how an attacker can steal personal information from existing popular social networking sites. They use two types of attacks to achieve this. The first one is based on the cloning of existing user accounts and the automated sending of friend requests, trying to trick the contacts of the cloned victim. The second one is more sophisticated and uses cross-site cloning of a user's profile and the containing contacts that exist in a social network, to another social network where the specified user has not yet registered.

3. CHALLENGES OF SOCIAL NETWORKS

New technologies lead to new challenges. The massive adoption of online social networks by hundreds of millions of users around the world has led to the emergence of many challenges. In this section we present certain aspects of a *fundamental challenge* posed by online social networks; the *public availability of personal information* that compromises users' privacy.

Social networks are one of the most popular and time consuming online activities with an average Facebook user spending more than 55 minutes a day on the site [3]. With users being attracted to OSNs, among other, for the ability to "socialize" with a large, geographically dispersed set of friends, as well as meet new people, users tend to befriend a much larger set of people than they would in the real world. With an average Facebook user having 130 online friends, many of whom are merely "cyber-acquaintances" [29] and posting a plethora of personal information that all

of them can access, social networks are leading to the age of unprecedented public availability of personal information. However, users do not comprehend the dangers of revealing personal information to online buddies many of whom they have never met in the real world [19]. While social networks provide security mechanisms to block access to certain personal information, studies have revealed that users do not comprehend issues of online privacy. Therefore, a challenge is to educate users on matters of online privacy so as to comprehend that the exposure of sensitive information is potentially dangerous.

However, we identify the challenge that arises from the participation in a social network, in regards to being targeted by attackers. We believe that the visibility of a user's participation in a network may offer enough information to attackers to make him the target of sophisticated personalized attacks. No matter how strict privacy settings may be introduced in the future, the names of almost all users will always be available to everyone. This is enough information for an attacker to use with a search engine and harvest email addresses faster and more efficiently than traditional harvesting techniques and, as shown in Section 5.2, map them to the owners' names. Even though Facebook users can use the security settings to prevent their profile from appearing in search results, few users will use it. A social network where users cannot find other users is, by nature, not viable. Therefore, while names must be visible to all, their automatic extraction must be hindered, as we propose in Section 7. A recent press release by Facebook urges users to make their posts public [18], which may lead to the public availability of even more personal information.

Default settings for Facebook and Twitter allow everyone to view a user's name, friends and pages he is a fan of. A study conducted by Gross et al [21] revealed that only 0.06% of the users hide the visibility of information such as interests and relationships, while in [23] the authors report that 99% of the Twitter users that they checked retained the default privacy settings. Attackers that harvest this publicly available information can use it to craft personalized attacks that are far more effective than traditional attacks. In section 7 we discuss several measures that can be employed to hinder the harvesting of personal information from on line social networks.

4. HARVESTING EMAIL ADDRESSES

In this section we give a brief overview of the current methodologies used by spammers to harvest email addresses.

Web crawling. Email addresses of users are posted in various places on the Web. Personal web pages, blogs and forums are such examples. By crawling the web attackers can gather thousands of email addresses. However, this methodology suffers from low scalability as web crawling is a very time-consuming and bandwidth-demanding process.

Crawling archive sites. Attackers can narrow down their crawling to sites they know contain thousands of email addresses. For example, the Mailing List Archives site [9] hosts archives for thousands of computer-related mailing lists. The obfuscation used to prevent crawlers from extracting addresses is very simple to bypass, as addresses are written in the form "username () domain ! top-level-domain".

Malware. Attackers can instrument their malware code to collect addresses from the email clients of infected users or their instant messaging clients. Given the widespread use of email clients and popularity of instant messaging networks, this technique provides good scalability.

Malicious sites. Attackers can lure users to sites and request for their email addresses in exchange for providing porn content and warez sites can offer access to movies and software provided that the user "registers" with their service.

Dictionary attacks. One can form email addresses by taking words from a dictionary. For example, the spammer can concatenate the word "john" with the domain "hotmail.com" and form the email address john@hotmail.com. Dictionary attacks can be classified into one of two types: blind attacks and search-based ones. Blind attacks try to guess email addresses by random concatenation of dictionary words and popular email domains. In this case, the attacker would send spam to "john@hotmail.com" without any knowledge of the validity of the email address. This approach is not efficient and is limited to the dictionary size. Search-based attacks make use of Web search engines to validate the addresses acquired by the dictionary concatenation. The attacker now searches for "john@hotmail.com" and parses the results for email addresses. This approach is more efficient as it can return more addresses than expected. As an example, searching for "john@hotmail.com" can also lead to "other.john@hotmail.com" and "john@hotmail.de".

In this work we describe a new approach on how attackers can use information from social networks to perform more advanced search-based dictionary attacks. Instead of using words from a dictionary, an attacker can crawl popular social networks and use the collected user names or pseudonames as search keywords. This approach has two major advantages. First, it scales with the growth rate of social networks. While dictionaries are limited to few hundred thousand terms, the number of user names and pseudonames that can be found in social networks is in the order of hundreds of millions. Second, information from social networks can be used for personalizing spam campaigns. For example, attackers can use the full names of users in order to construct more convincing spam emails. We describe our approach in more detail in Section 5.

5. USING SOCIAL NETWORKS TO HARVEST EMAIL ADDRESSES

Social networks provide a plethora of personal information. Users upload reports from their daily activities, political and religious status, events they have or will attend, photos, comments for other users and many more. Once a user has managed to become a friend with someone, he can extract various pieces of information that can be used for illegal purposes.

Even though social networking sites cannot protect users from other malicious users that want to harvest personal information through social engineering tricks, they protect email addresses from automated harvesting. Before we describe how to use social networks as harvesting engines, we present the defensive measures taken by two popular social networking sites, Facebook and Twitter. Facebook does not reveal a user's email address to any other user that is not in his friend list. In case the harvester is in the list, the user's email address is presented as a GIF image to prevent automated extraction. Twitter, on the other hand, does not reveal a user's email address in any form. However, the per-

sonal information that is revealed includes the user's name, personal web page, location and a short bio description.

We identify and outline two different strategies that spammers may follow depending on the type of spam campaigns they wish to promote. First, we have spammers that propagate emails that contain advertisements for various products. This type of spammer will follow the *blind harvesting* approach which is the technique that will result in gathering as many email addresses as possible. Second, we have spammers that use spam emails to propagate scams, such as phishing campaigns. This type of spammer will use the *targeted harvesting* technique that returns a much smaller number of results, but harvests information that can be used to craft very convincing personalized emails.

5.1 Blind harvesting

This technique aims to blindly harvest as many email addresses as possible in an efficient manner. The spammer does not care for personal information but simply wishes to gather email addresses. As shown by our results in Section 6.1, using social networks in conjunction with search engines is the most efficient method to harvest large numbers of email addresses.

We follow the same approach for both Facebook and Twitter to harvest email addresses. We initially crawl both networks to find names. As the structure and properties of the Facebook and Twitter networks differ, we have implemented two different crawlers for extracting names. One might use the Facebook search utility to search for and harvest names. However a far more efficient way is to use Facebook fan pages. Users become fans of an artist or an activity. One can freely browse all the names of a fan page. For example, the fan pages of Madonna and Shakira (popular pop artists) have 1.3 and 1.7 million fans respectively, while Barack Obama has 8.8 million. Any attacker can visit a popular fan page, and will immediately have access to millions of names. In the case of Twitter we started from one initial account and then crawled the accounts the user follows, then the accounts they follow and so on. As we were interested only in the users' names and nicknames and not the actual tweets, this simple crawling is effective and fast for harvesting names.

Once the names have been harvested, they are used as terms in a search engine query. We used the Google search engine to locate email addresses. For each search term we query 8 different combinations ("term@hotmail.com", "term", "term@msn.com", "term@windowslive.com", "term@", "term at", "term@gmail.com", "term@yahoo.com") and for each query we retrieve the first 50 results. For scalability and efficiency reasons we do not open the URLs returned by the search engine. Instead, we parse the two-line summary provided in the results, for email addresses. This results in us missing a number of email addresses that may not be returned in the summary, however we remove a large overhead of having to parse the whole page. Our parser takes into account the various techniques used to hide email addresses from web crawlers, such as "username [at] domain".

5.2 Targeted harvesting

Attackers that rely on spam messages to propagate phishing schemes, can craft personalized phishing emails that are far more efficient than traditional techniques, by using personal information publicly available in social networks. Even though the blind harvesting technique can collect millions of email addresses efficiently, it presents a low probability of having these addresses matched to the name of their owners. The targeted harvesting approach links names to email addresses with a high probability, if not, absolute certainty. Furthermore, it enables the gathering of additional information that can render a targeted message much more convincing. Depending on the attack and the amount of personal information the attacker wants to collect, we describe three different methodologies for targeted harvesting.

Reverse lookup emails on Facebook. In the first case, we rely solely on the email-based search functionality of Facebook. Facebook allows users to search for other users based on their email address. We were surprised to find that even if the user has protected his email address through the privacy settings, and has made it visible only to him, his name will still show up in the search results when someone searches his email address. Only if the user disables his inclusion in public search results, we will not be able to find him using his email address. However, by default, Facebook includes users in search results. We collect names from highly populated Facebook fan pages and use the blind harvesting technique to search for email addresses using Google. We then search for the harvested email addresses in Facebook and obtain the results. This way we have a pair of a user's profile and his email address (and any other information that is public), the basic information needed for a personalized phishing email. We can augment the collected information of the matched users by inviting them to become our friends. Once a user has accepted, we now have access to all the information posted in his Facebook profile. Our results from a series of initial experiments showed that 30% of the random invitations were accepted.

A major advantage of this technique is that it not only maps an email address to the owner's social profile, but also provides a technique for validating email addresses without the need of sending "probing" emails. When no profile is returned for a specific email address we cannot conclude if the email address is valid or not. However, when a user's profile is returned, we ascertain that the specific email is valid, since the user has entered it in his profile's contact information. Therefore, all the email addresses harvested using this technique are valid and eliminate the overhead of sending spam emails to many email addresses that are not valid. This is another advantage for spammers, since by eliminating all the emails that would be sent to invalid addresses and reducing the overall volume of the spam emails they send, they may be able to evade spam detection systems [30] that rely on the collection of a large number of spam emails.

Nickname-based Email Harvesting. In the second case we aim to use information that is available on Twitter in order to narrow down the search space of our first technique and improve its efficiency. This is done by using the nickname information available on Twitter. Many people tend to create a nickname that they consistently use across different domains and email providers. Our method crawls Twitter and collects name and nickname pairs. We then query Google using the nickname as a search term and extract email addresses that are an exact match (for example, if the nickname was "john_doe_1" we would only extract emails of the form "john_doe_1@domain.com"). This provides an association between a name and one or more email addresses. Next, we use the harvested email addresses as

terms in the email-based search functionality of Facebook, exactly as in the first technique. Using this approach, one has to check much fewer email addresses than the first technique and, additionally, the success rate is higher as Twitter users will probably also have a Facebook account. The innovation of this technique is that it combines disjoint sets of personal information publicly available on different social networks and can be fully automated.

Site-aware Harvesting. In the third case, we employ Google's Buzz [5], a recently launched social networking service. In a nutshell, Buzz is a Twitter-like social networking service (based on follower/followee relations), along with content feeds and integration with other Google services (Gmail, Google Reader, Picassa, YouTube etc.). Each Buzz user has a Google profile page that contains basic information about him and his follower/followee relations. The Google profile page URL can either be based on the Google account username or a random long numeric identifier. The Google account username acts as a global identifier for all Google services, including the Gmail service. This means that if a user's Google profile URL includes his username and the user appears in the Buzz graph, then we automatically know his Gmail address. Thus, we can use the social graph of Buzz as a means to discover Gmail addresses. This approach has two major advantages. First, all harvested emails are valid. Second, and most important, for all collected email addresses we have the name of their owner, as we can extract it from the corresponding profile page. Moreover, since Buzz actually prompts the user to link and fetch content from other sites such as Twitter, Flickr, Google Reader, YouTube, FriendFeed and LinkedIn, the attacker can enrich the amount and type of information assembled and utilized for the targeted spam campaign. We crawl Buzz profiles, through the Buzz search feature, by looking up names collected from Facebook and extract the follower/followee relations, wherever it is feasible. Additionally, references to unrelated profiles are returned by the search results as part of the indexed content. In the case where the user hides his relations, we are still able to process the profile contents, comprised of messages from and to other users. All names, that are rendered as clickable links to their respective profile pages, have their profile identifiers exposed. Even if Buzz decides to remove these links, effectively crippling the usability of the profile page, we could simply collect their names and look them up separately through the Buzz search feature.

6. MEASUREMENTS

Here we evaluate the proposed email harvesting techniques described in detail in Section 5. Furthermore we compare our techniques with the currently used approaches described in Section 4. Finally we perform a study regarding the use of harvested information in a spam campaign.

6.1 Blind Harvesting

We evaluate the use of our blind harvesting technique in comparison to current approaches. For obvious reasons we have omitted the malware and malicious site approaches from our comparison. Before proceeding to the analysis we first present and explain the comparison axes of our evaluation. We use three metrics:

- **Addresses-per-keyword ratio.** It is one of the most important metrics. A low ratio means that for each keyword queried the number of email addresses harvested is low. A high ratio means that the methodology can extract tens or hundreds of email addresses per keyword.

- **Traffic volume ratio.** Using search engines and sites for harvesting purposes requires downloading millions of pages. Downloading Gigabytes of data to harvest only a few email addresses decreases the scalability of the approach.

- **Automation.** Harvesting methodologies must be automated in order to be efficient. Although some approaches present high addresses-per-keyword ratio, they require manual intervention as they use information that does not expand and is located in multiple locations.

	Dataset	Unique emails	Ratio
Facebook Names	82,383	3,706,493	1:45
Twitter Names	87,334	2,012,391	1:23
Twitter Nicks	31,358	784,099	1:25
Dictionary	146,973	3,630,071	1:24.7
Surnames	23,300	2,200,225	1:94
Documents	680,973	445,678	1:0.65
MARC	438,722	5,265	1:0.012
W3C	376,641	330,436	1:0.87

Table 1: A detailed listing of the dataset size and the number of unique email addresses harvested for each technique.

Address-per-keyword Ratio. Our first measurement evaluated the addresses-per-keyword ratio between our blind harvesting technique and four traditional harvesting methods: crawling archive sites, crawling the web for documents, a generic dictionary attack and a specialized dictionary attack. We crawled the MARC [9] and the W3C archive [17] sites to search for email addresses. For the document harvesting experiment, we only retrieved MS Word, Excel, Powerpoint and PDF documents as a step to narrow down our search space. For the generic dictionary attacks, we used keywords from an English dictionary [8]. For the specialized dictionary attack we used the 23,300 most popular English surnames [4]. For our harvesting techniques we extracted user names from Facebook and Twitter as well as user "nicknames" from Twitter. In all the experiments, we extracted all email addresses from the Google query results, and additionally evaluated the case where email addresses were an exact match to the Twitter nicknames.

The results are summarized in Figure 1. In the case of Facebook we extracted emails with a ratio of 1:45, i.e., we were able to harvest, on average, 45 unique email addresses per name queried. Using Twitter names, we achieved a ratio of 1:23, while a dataset of nicknames returned 25 addresses per query. The highest ratio observed was by the specialized version of the dictionary attack, which yielded 94 addresses per keyword. In fact, this methodology was expected to harvest a larger number, as it follows a similar approach but takes the most popular English names. However, this method suffers from scalability issues as described later in this section. The generic dictionary attack, contrary to the

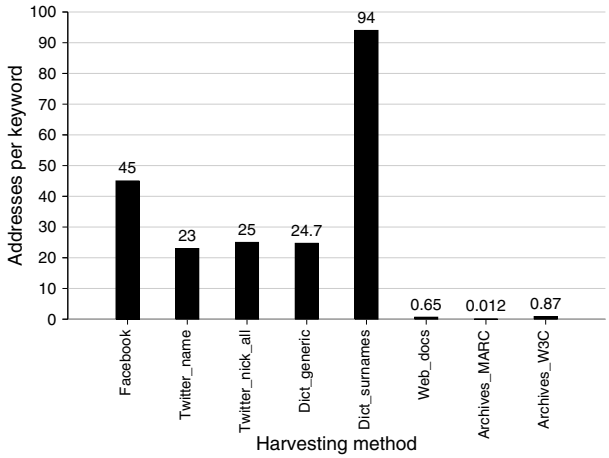

Figure 1: Ratio of unique email addresses per keyword for various email harvesting methodologies.

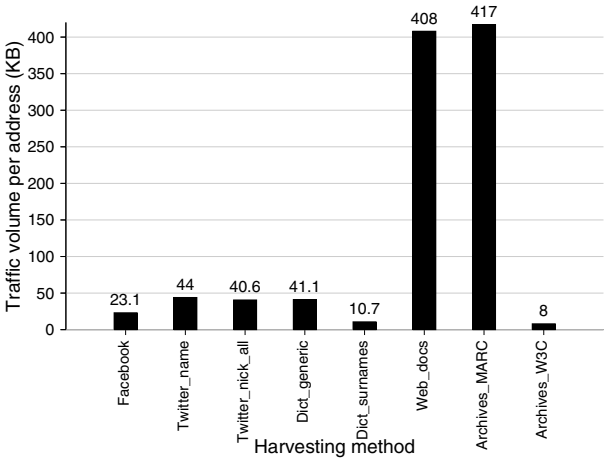

Figure 2: Ratio of traffic volume per email address for various harvesting methodologies

specialized one, achieved a lower ratio of 1:24.7. Crawling the web for documents returned 0.65 addresses per file downloaded. Finally, in the case of archive site crawling, the ratio for MARC and W3C archives is 1:0.012 and 1:0.87 respectively, where the ratio is defined as addresses extracted per page fetched. The low ratio for crawling sites is due to the download of structure pages, which are pages without any email address that contain hyperlinks to pages deeper in the site hierarchy. In fact, 96.7% of the MARC pages were structure pages as this site is deeply nested. The W3C archive follows a more flat structure: 16.5% of the pages were structure pages. Ideally, if we exclude the structure pages, the ratios for the MARC and W3C archive become 1:0.4 and 1:1.05 respectively. Table 1 depicts the size of the aforementioned datasets, along with the count of harvested email addresses which produce the respected ratios.

Traffic Volume Ratio. Our second metric focuses on the cost per email address in Kbytes. The results are summarized in Figure 2. The traffic volume for the Facebook case is the number of names times the page size of Google results, that is 82,383 names times 130 Kbytes per Google result page times 8 (8 search combinations per name). The total traffic volume is around 79.8 Gbytes for approximately 23.1 Kbytes per email address. In the case we use names taken from Twitter, the ratio is 44 Kbytes per email address. When we use Twitter nicknames, the ratio drops down to 40.6 Kbytes per address. In the case of downloading office documents, the total volume of files was 181.6 Gbytes plus an additional 1.6 Gbytes for the Google queries, that is 408 Kbytes per email address. For the generic dictionary attack, we retrieved 142.3 Gbytes of search results which gives a ratio of 41.1 Kbytes per email address. For the specialized dictionary attack using popular surnames, we fetched 22.5 Gbytes of search results, that is a ratio of 10.7 Kbytes per email address. Finally, for the archive site crawling experiments, we downloaded 4.6 Gbytes, a ratio of 14.8 Kbytes per address in total. If we examine the two archive sites separately, the ratio for MARC is 417 Kbytes per address and for W3C is 8 Kbytes per address.

Automation. Our proposed harvesting technique is highly scalable. As we use information retrieved from social networks, our approach follows their growth rate. Therefore, our technique is fully automated as it expands, and no further manual intervention is needed for collecting more names that will be used as seeds. On the other hand, document crawling, generic dictionary attacks, and attacks based on surnames present very low scalability as the search terms are static, unlikely to change and have a limited dictionary size. Therefore, the process is semi-automated as customized crawlers have to be implemented for all new sites incorporated. Crawling mailing list archives presents medium scalability as we extract information from communities that expand, but that are interested in specific topics and expand with a much slower rate than social networks. This technique is also a semi-automated process, as most of the sites follow their own format to depict email addresses, and the appropriate regular expressions have to be written by hand.

Overall, while the harvesting technique that uses surnames presents a higher ratio for keywords per email and a smaller cost, it is not the optimal and most efficient one as it relies on a finite and limited dictionary that does not expand. On the other hand, while the *blind harvesting* technique exhibits a lower ratio and slightly higher cost, it has the advantage of being scalable, as it follows the expansion rate of social networks. *In the long run, we consider this to be the optimal solution for large-scale efficient harvesting.*

6.2 Effectiveness of Targeted Harvesting

The second part of our evaluation focuses on our targeted harvesting techniques. Our experiment aims at measuring the effectiveness of these techniques for conducting personalized phishing campaigns. The results depict the percentage of names for which we can harvest at least one of their actual email addresses with each technique and therefore represent its effectiveness. We created two datasets containing randomly selected names from our databases. For reasons explained below, we selected names comprised solely of a first and last name, excluding middle names, dots or hyphens.

The first dataset contained 9000 names collected from a Facebook fan page. We used this dataset to evaluate our **first targeted harvesting technique**: for each name, we blindly harvested email addresses using the name as a search term in the Google engine and collected **any search re-**

sults. We then looked up the harvested email addresses using the Facebook search feature. If one or more profiles were returned, we checked whether any of them had a matching name with the one collected from the Facebook fan page and coupled with the email address in question. Overall, about 11.5% of unique names were associated with an email address that yielded a matching profile result from Facebook.

The second dataset was collected from crawling the Twitter network. For the **second targeted harvested technique** we wanted to measure the effectiveness of employing strict heuristics during the initial collection of email address through Google Search. For that matter, we included only **exact match results** of email addresses, i.e. only those whose prefix was identical to the Twitter username of the user being queried. Overall, using this strict Google search heuristic, we assembled 38986 <name,email> tuples, corresponding to 15627 unique names collected by our Twitter crawler. From those names, we selected 8,986 which did not contain middle names or special characters, just like in the first experiment. The reason for this filtering lies on the straightforward verification heuristic we employed; for each email address coupled with a name, we looked it up using Facebook search and, from any profile results returned, considered a match only if the name was exactly the same as the one in the dataset. Therefore, entries with middle names or special characters, having a larger possibility of being written differently across disjoined social networks, were excluded. The addresses were grouped by the Twitter nickname that resulted in their discovery. From the 8,986 users, 3,588 (39.9%) returned a Facebook profile and 1,558 (17.8%) were an exact match. Thus, 43.4% of the names, that returned a profile, had a user name that was an exact match to the Twitter profile name. By using a fuzzy string matching approach we could improve the success percentage. Let there be noted that names, that their harvested emails did not yield any Facebook results, may or may not be true positives of the targeted harvesting technique. As discussed in Section 8, additional OSNs could be employed to improve the query dataset. Also, in section 6.3 we present a study regarding the personal info collected from these profiles.

In comparison, the first and second methodologies, i.e., loose and strict collection of email address from Google search, may appear to be similarly effective with 11.5% and 17.8% of the names being a match. However, in the first case, a name is coupled with a much greater set of possible email addresses, requiring far more lookups in the Facebook than the second. In detail, in the first case, each name was coupled with an average of 104 email address, while, in the second case, only 4 address lookups took place for each name. Consequently, in the first case 0.2% of email address returned a profile result with a matching name, while in the second case the effectiveness climbed to 7%.

In regards to the **Google Buzz** approach, we used 1705 names and 850 of the most common English words (such as book, chair etc.) as search terms. We gathered a total of 59,680 Google profile URLs. 40.5% of the Google profile URLs (24,206 profiles) included the users' Google username, also used by default as their email address prefix, while the rest of the profiles were assigned random identifiers. This means that for each search term we gather approximately 22 Google profile URLs and around 9 valid Gmail accounts. As mentioned in section 5, all email addresses extracted from the profile usernames are valid Gmail accounts.

Label	Popularity
Current City	41.8% (667)
Hometown	38.8% (619)
Employers	24.9% (397)
College	24.5% (391)
High School	24.1% (385)
Relationship Status	21.0% (335)
Grad School	8.8% (140)
Birthday	3.9% (63)
Anniversary	3.4% (54)
Religious Views	2.5% (40)
Political Views	2.3% (36)

Table 2: Selected labels of personal information available on a Facebook profile page and their respective popularity among the matching profiles of the targeted harvesting evaluation.

Category	Frequency
TV/Cinema	50%
Music	24%
Activity/Sports	10%
City/Travel	11%
Various	3%
Technology	2%

Table 3: Content categorization of the 100 most frequent items in a Facebook profile page.

6.3 Study of harvested personal info

In this section we present a study based on the personal information publicly available in the Facebook profiles harvested from our second *targeted harvesting* technique. As mentioned in Section 6.2, 1,558 unique names were associated with a least one email address which yielded an exact-math profile match in Facebook, thus verifying the initial <name,email> association made by the Twitter crawler. Some of those names had more than one email addresses providing matching profiles. We investigated those cases and concluded that the profiles belonged to different people that shared the same name. Overall, 1,558 names led to 1,597 distinct profiles.

In Table 2 we present some selected labels of information, availabe on the Facebook profiles we harvested, which we consider to reveal personal information that can be exploited by attackers for targeted phishing attacks. For instance, one may use information about current employers or a person's studies to fake a workplace or college-related message. By adding such information, the email becomes more convincing and is therefore more likely to fool its recipient. For a full list of the categories, the reader may refer to the Appendix, at the end of this paper.

Subsequently, we proceed to examine the content of the Facebook profile, i.e., the page elements. We select the top 100 that appear more frequently among our dataset and apply a manual categorization. Table 3 summarizes the results. One may observe that items related to TV and cinema are the most common. An attacker could lure victims by crafting phishing messages to include references to such popular content.

As shown by recent phishing campaigns [7], attackers use information regarding a victim's Facebook contacts, to impersonate their friends and trick them into giving them money. This type of attack could easily propagate to email phishing

campaigns. To measure the feasibility of such attack, we calculate the percent of the harvested profiles which expore their respective friend lists. Overall, 72.6% of them, leak such information and the mean number of friends is 238.

7. DISCUSSION

In this section we provide a discussion on various measures that can minimize the public availability of personal information and hinder attackers from easily harvesting such information. While the defenses proposed can enforce users' privacy, we also refer to their potential negative impact on the functionality and expansion of social networking sites.

Server-Side Security. When proposing these measure, one must take into consideration that users of social networking sites are not restricted to "computer-savvy" people. In fact, the accessibility of such sites through mobile phones, smartphones and handheld devices allows the participation of people who do not even own a computer. For that matter, we consider the familiarity of users with computers to be minimal and their knowledge regarding information security and privacy matters to be negligible. Overall, it is our belief that any privacy measures taken should lie on the server-side and, therefore, propose only such.

Strict Privacy by Default. The first step that needs to be taken by social networking sites is to enforce strict default privacy settings. As shown by previous work [23], most users do not change default privacy settings and, thus, expose a large amount of information. For instance, Facebook's default settings reveal a person's real name, photograph, sex, relationship status, gender preferences, current city, hometown, biography, favorite quotations, current and previous employers, college and high school education, interests in music, books, movies and television and personal website. The e-mail address is not exposed but by searching for it, the person's profile will be returned. One should not be able to view any information from a user's profile other than his name if they are not friends in the specific networking site. If OSNs opt to hide all user information from third parties, attackers will not be able to harvest information for crafting personalized phishing attacks. On the other hand, features, such as email-address-based profile search, provide the necessary functionality for the social network to expand. Upon registration, a new user may use this feature to identify which of his e-mail contacts exist in the network and therefore instantly boost his networking degree.

Information-leakage Indicators. A variation of the first step is the preservation of standard privacy settings and the addition of indicators (e.g. icons, colors), that only the user can see, next to each profile field, illustrating information that is publicly available (e.g. any Internet user has access to it - colored red), available within the network (e.g. any Facebook user has access to it - colored orange), available within friends (e.g. any friend/contact of the user has access to it - colored yellow) and available only to the user himself (colored green). We believe that users will be very receptive to this concept as they will be able to instantly identify, through a glance at their profile, information that is exposed despite their will or knowledge. Nonetheless, social networking sites operate with the need for users to provide as much information possible about themselves. Such privacy indicators could scare the user into withdrawing a substantial amount of information.

Information rendered as Images. The next measure that can hinder attackers from harvesting names that can lead to email addresses, is to display names as images, just like the way Facebook presents e-mail addresses. Displaying names as images raises the difficulty for extracting them, increases the error ratio on the attacker's side and does not break the users' experience. However, social networks can also provide a way for displaying names as plain text after verifying that the entity that issued the request is not a bot, e.g., by using CAPTCHAs. Unfortunately CAPTCHAs are not fool-proof. For instance, in [20], the authors were able to solve social networking site CAPTCHAs, including Facebook's reCAPTCHAs, through simple image processing techniques, combined with a dictionary and Google searches.

Automatic Tools Detection. Furthermore, OSNs should employ techniques that can detect accounts that are used by bots either to automatically issue friend requests for harvesting purposes or flood users with spam advertisements. Several services [1] are available and one is already being used by Twitter [14]. We believe this to be a major step in protecting users' privacy, since a large fraction of users accepts friend requests from unknown profiles. Therefore, all social networking sites must deploy such services.

Email Reverse Search. A major blunder on the side of social networking sites, is to allow users to search for profiles by using email addresses. By doing so, an attacker can easily map harvested email accounts to user profiles, and use the publicly available information to craft very convincing personalized phishing emails.

Use of nicknames. We believe that OSNs should exhibit the following behavior regarding the use of nicknames: if a user is logged in the site and is also in the contact list of the person using a nickname, he should be able to use the nickname directly (e.g., `facebook.com/nickname`). In any other case, the OSNs will prohibit its use, returning a "nickname not found" error. Instead of the nickname, a unique and random identifier will be used (e.g., `facebook.com/1309501319510`). This way, another user coming across this profile reference (e.g., in a fan page) will be unable to obtain the actual nickname and map it to an email address.

8. FUTURE WORK

The targeted harvesting technique relies on online social networks to verify associated pairs of names and email addresses and, subsequently, gather additional information about their owners. In this paper we have employed Facebook for that purpose. However, a plethora of social networks exists. In fact, search engines such as pipl [11], spokeo [15] and knowem[6] provide aggregated search results across hundreds of user networks. One could make use of these services to perform more extensive lookups, thereby improving the efficiency of the technique. Specifically, in the cases where email lookups in Facebook return no results or irrelevant profiles these services could provide meaningful info from other online social networks.

9. CONCLUSIONS

In this paper we present how information, that is publicly available in social networking sites, can be used for harvesting email addresses and deploying personalized phishing campaigns. We argue that an inherent challenge of a social network is the visibility of its members. The mere participation of a user renders him a target for personalized

attacks. We present two different approaches to harvesting email addresses. *Blind harvesting* uses names collected from social networking sites and aims to collect as many email addresses as possible. Using this technique we were able to harvest millions of email addresses in an efficient fashion. *Targeted harvesting* aims to harvest email addresses that can be mapped to a name and publicly available information and, thus, greatly enhance the efficiency of a spam campaign. We present three such techniques. The first technique blindly harvests email addresses and uses Facebook to map them to a user name, with a success rate of 11.5%. By using information available in the Twitter network we are able to narrow the search space and accurately map 43.4% of the user profiles. Next, we use names collected from Facebook fan pages to harvest Google Buzz accounts, 40.5% of whom provide a direct mapping to a Gmail account. Finally, we present a discusssion of various defense techniques that can hinder attackers from using online social networks to harvest email addresses and personal information.

Acknowledgments

This work was supported in part by the project SysSec funded in part by the European Commission, under Grant Agreement Number 257007. We thank the anonymous reviewers for their valuable comments. Iasonas Polakis, Georgios Kontaxis, Eleni Gessiou, Thanasis Petsas and Evangelos P. Markatos are also with the University of Crete.

10. REFERENCES

[1] Clean Tweets. http://blvdstatus.com/clean-tweets.html.

[2] Facebook entry gets office worker fired. http://news.cnet.com/8301-17852_3-10172931-71.html.

[3] Facebook statistics. http://www.facebook.com/press/info.php?statistics.

[4] Genealogy data: Frequently occurring surnames. http://www.census.gov/genealogy/www/data/2000surnames/index.html.

[5] Google buzz. http://buzz.google.com/.

[6] KnowEm. http://knowem.com/.

[7] Latest Facebook Scam: Phishers Hit Up "Friends" for Cash. http://techcrunch.com/2009/01/20/latest-facebook-scam-phishers-hit-up-friends-for-cash/.

[8] A lexical database for English. http://wordnet.princeton.edu/.

[9] Mailing list archives. http://marc.info/.

[10] More employers use social networks to check out applicants. http://bits.blogs.nytimes.com/2009/08/20/more-employers-use-social-networks-to-check-out-applicants/.

[11] pipl - search people of the web. http://pipl.com/.

[12] Please Rob Me. http://pleaserobme.com/.

[13] Project honey pot: Help stop spammers before they ever get your address! http://www.projecthoneypot.org/.

[14] Spam detector. http://www.cs.ucsb.edu/~gianluca/spamdetector.html.

[15] spokeo. http://www.spokeo.com/email.

[16] Twitter blog, measuring tweets. http://blog.twitter.com/2010/02/measuring-tweets.html.

[17] W3C public mailing list archives. http://lists.w3.org/.

[18] Facebook blog: More ways to share in the publisher. http://blog.facebook.com/blog.php?post=98499677130, 2006.

[19] Acquisti, A., and Gross, R. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *Proceedings of 6th Workshop on Privacy Enhancing Technologies* (2006).

[20] Bilge, L., Strufe, T., Balzarotti, D., and Kirda, E. All your contacts are belong to us: automated identity theft attacks on social networks. In *WWW '09: Proceedings of the 18th international conference on World wide web* (2009), ACM, pp. 551–560.

[21] Gross, R., Acquisti, A., and Heinz, III, H. J. Information revelation and privacy in online social networks. In *WPES '05: Proceedings of the 2005 ACM workshop on Privacy in the electronic society* (New York, NY, USA, 2005), ACM, pp. 71–80.

[22] Kreibich, C., Kanich, C., Br, K. L., Enright, O., and Savage, S. On the spam campaign trail. In *In First USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEETâĂŹ08)* (2008).

[23] Krishnamurthy, B., Gill, P., and Arlitt, M. A few chirps about twitter. In *WOSN '08: Proceedings of the first workshop on Online social networks* (New York, NY, USA, 2008), ACM, pp. 19–24.

[24] Krishnamurthy, B., and Wills, C. E. On the leakage of personally identifiable information via online social networks. In *WOSN '09: Proceedings of the 2nd ACM workshop on Online social networks* (New York, NY, USA, 2009), ACM, pp. 7–12.

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

[26] Shue, C. A., Gupta, M., Lubia, J. J., Kong, C. H., , and Yuksel, A. Spamology: A study of spam origins. In *The 6th Conference on Email and Anti-Spam (CEAS)* (2009).

[27] Smith, A. Nigerian scam e-mails and the charms of capital. *Cultural Studies 23*, 1, 27–47.

[28] Tang, J., Musolesi, M., Mascolo, C., and Latora, V. Temporal distance metrics for social network analysis. In *WOSN '09: Proceedings of the 2nd ACM workshop on Online social networks* (2009), ACM, pp. 31–36.

[29] Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P. On the evolution of user interaction in facebook. In *WOSN '09: Proceedings of the 2nd ACM workshop on Online social networks* (2009), ACM, pp. 37–42.

[30] Xie, Y., Yu, F., Achan, K., Panigrahy, R., and Hulten, G. Spamming botnets: signatures and characteristics. In *In SIGCOMM* (2008).

[31] Xu, W., Zhou, X., and Li, L. Inferring privacy information via social relations. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on* (2008), pp. 525–530.

APPENDIX

A. PERSONAL INFO

Label	Popularity
Sex	70.1% (1119)
Facebook Profile	59.9% (957)
Music	46.9% (749)
Current City	41.8% (667)
Hometown	38.8% (619)
Television	38.1% (609)
Movies	33.8% (539)
Website	29.2% (466)
Employers	24.9% (397)
College	24.5% (391)
High School	24.1% (385)
Interests	21.1% (337)
Activities	21.1% (337)
Relationship Status	21.0% (335)
Books	18.9% (302)
Favorite Quotations	13.8% (220)
Other	13.7% (218)
Bio	13.6% (217)
Looking For	12.5% (200)
Grad School	8.8% (140)
Interested In	7.3% (116)
Siblings	6.4% (102)
Birthday	3.9% (63)
Parents	3.4% (55)
Children	3.4% (55)
Anniversary	3.4% (54)
Email	2.9% (46)
Religious Views	2.5% (40)
Political Views	2.3% (36)
AIM	0.5% (8)
Mobile Number	0.4% (7)
Skype	0.4% (6)
Address	0.4% (6)
Google Talk	0.3% (5)
Windows Live	0.2% (3)
Yahoo	0.1% (2)
Phone	0.1% (1)

Table 4: Labels of personal information available on a Facebook profile page and their respective popularity among the matching profiles of the targeted harvesting evaluation.