Reverse Social Engineering Attacks in Online Social Networks

Conference Paper · July 2011

DOI: 10.1007/978-3-642-22424-9_4 · Source: DBLP

5 authors, including:

Davide Balzarotti
EURECOM
122 PUBLICATIONS **6,011** CITATIONS

SEE PROFILE

Calton Pu
Georgia Institute of Technology
359 PUBLICATIONS **7,645** CITATIONS

SEE PROFILE

Reverse Social Engineering Attacks in Online Social Networks

Danesh Irani[1], Marco Balduzzi[2], Davide Balzarotti[2]
Engin Kirda[3], and Calton Pu[1]

[1] College of Computing, Georgia Institute of Technology, Atlanta
[2] Institute Eurecom, Sophia Antipolis
[3] Northeastern University, Boston

Abstract. Social networks are some of the largest and fastest growing online services today. Facebook, for example, has been ranked as the second most visited site on the Internet, and has been reporting growth rates as high as 3% per week. One of the key features of social networks is the support they provide for finding new friends. For example, social network sites may try to automatically identify which users know each other in order to propose friendship recommendations.
Clearly, most social network sites are critical with respect to user's security and privacy due to the large amount of information available on them, as well as their very large user base. Previous research has shown that users of online social networks tend to exhibit a higher degree of trust in friend requests and messages sent by other users. Even though the problem of unsolicited messages in social networks (i.e., spam) has already been studied in detail, to date, reverse social engineering attacks in social networks have not received any attention. In a reverse social engineering attack, the attacker does not initiate contact with the victim. Rather, the victim is tricked into contacting the attacker herself. As a result, a high degree of trust is established between the victim and the attacker as the victim is the entity that established the relationship.
In this paper, we present the first user study on reverse social engineering attacks in social networks. That is, we discuss and show how attackers, in practice, can abuse some of the friend-finding features that online social networks provide with the aim of launching reverse social engineering attacks. Our results demonstrate that reverse social engineering attacks are feasible and effective in practice.

Keywords: social engineering, social networks, privacy

1 Introduction

Social networking sites such as Facebook, LinkedIn, and Twitter are arguably the fastest growing web-based online services today. Facebook, for example, has been reporting growth rates as high as 3% per week, with more than 400 million registered users as of March 2010 [2]. Many users appreciate social networks

because they make it easier to meet new people, find old friends, and share multimedia artifacts such as videos and photographs.

One of the key features of social networks is the support they provide for finding new friends. For example, a typical technique consists of automatically identifying common friends in cliques and then promoting new friendships with messages such as "*You have 4 mutual friends with John Doe. Would you like to add John Doe as a new friend?*". Also, information on the activities of users are often collected, analyzed, and correlated to determine the probability that two users may know each other. If a potential acquaintance is detected, a new friendship recommendation might be displayed by the social network site when the user logs in.

Clearly, social networks are critical applications with respect to the security and privacy of their users. In fact, the large amount of information published, and often publicly shared, on the user profiles is increasingly attracting the attention of attackers. Attacks on social networks are usually variants of traditional security threats (such as malware, worms, spam, and phishing). However, these attacks are carried out in a different context by leveraging the social networks as a new medium to reach the victims. Moreover, adversaries can take advantage of the trust relationships between "friends" in social networks to craft more convincing attacks by exploiting personal information gleaned from victims' pages.

Past research has shown that users of online social networks tend to exhibit a higher degree of trust in friend requests and messages sent by other users (e.g., [1, 5]). In addition, some forms of attacks on social networks, such as the problem of unsolicited messages, have already been studied in detail by the research community (e.g., [9, 16]). However, to date, *reverse social engineering* attacks in social networks have not received any attention. Hence, no previous work exists on the topic.

In a reverse social engineering attack, the attacker does not initiate contact with the victim. Rather, the victim is tricked into contacting the attacker herself. As a result, a high degree of trust is established between the victim and the attacker as the victim is the entity that first wanted to establish a relationship. Once a reverse social engineering attack is successful (i.e., the attacker has established a friend relationship with the victim), she can then launch a wide range of attacks such as persuading victims to click on malicious links, blackmailing, identity theft, and phishing.

This paper presents the first user study on how attackers can abuse some of the features provided by online social networks with the aim of launching automated reverse social engineering attacks. We present three novel attacks, namely, recommendation-based, visitor tracking-based, and demographics-based reverse social engineering. Furthermore, using the popular social networks Facebook, Badoo, and Friendster, we discuss and measure the effectiveness of these attacks, and we show which social networking features make such attacks feasible in practice.

In the recommendation attack, the aim is to exploit the friend recommendations made by the social network to promote the fake profile of a fictitious user

to the victim. The hope, from the attacker's point of view, is that the victim will be intrigued by the recommendation, and will attempt to contact the bogus profile that is under the attacker's control. In the visitor tracking attack, the aim is to trigger the target's curiosity by simply browsing her profile page. The notification that the page has been visited may be enough to attract the target to visit the attacker profile. Finally, in the demographic-based attack scenario, the attacker attempts to reach his victims by forging fake demographic or personal information with the aim of attracting the attention of users with similar preferences (e.g., similar musical tastes, similar interests, etc.).

Our findings suggest that, contrary to the common folk wisdom, only having an account with an attractive photograph may not be enough to recruit a high number of unsuspecting victims. Rather, the attacker needs to provide victims with a pretext and an incentive for establishing contact.

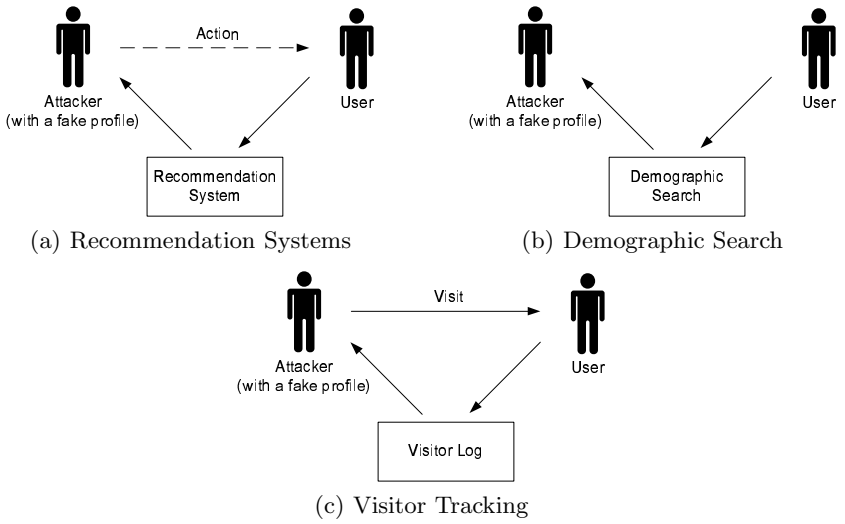
In this paper, we make the following contributions:

- We present the first user study on reverse social engineering in social networks and present three novel attacks. In particular, we discuss and measure how attackers can abuse some of the friend-finding features that online social networks provide with the aim of launching automated reverse social engineering attacks against victims.
- We measure how different user profile attributes and friend recommendation features affect the success of reverse social engineering attempts.
- We study the interactions of users with accounts that have been set up to perform reverse social engineering, and provide insights into why users fall victim to such attacks.
- We propose mitigation techniques to secure social networks against reverse social engineering attempts.

2 Reverse Social Engineering in Social Networks

Online social engineering attacks are easy to propagate, difficult to trace back to the attacker, and usually involves a low cost per targeted user. They are well-known threats in which the attacker aims at influencing the victims, and making them perform actions on her behalf. The attacker is typically interested in tricking the victims into revealing sensitive or important information. Examples of these attacks include traditional e-mail hoaxes and phishing, or their more advanced targeted forms, such as spear phishing.

Most online social engineering attacks rely on some form of "pretexting" [14]. That is, the attacker establishes contact with the target, and sends some initial request to bootstrap the attack. This approach, although effective because it can reach a large number of potential victims, has the downside that Internet users are becoming more and more suspicious about unsolicited contact requests. However, previous work has shown that it is possible to raise levels of trust by impersonating an existing friend of the target (e.g., [5, 10]) or by injecting the attack into existing chat conversations [13].



(a) Recommendation Systems (b) Demographic Search

(c) Visitor Tracking

Fig. 1. Different types of Reverse Social Engineering attacks.

Reverse Social Engineering (RSE) is a form of social engineering attack that has not yet been reported widely in an online context. RSE is a well-known technique in the hacker community (e.g., [14]) for targeted phone attacks. The attack, in a first step, relies on some form of "baiting" to stimulate the victim's curiosity. In a second step, once the victim's interest is raised, the attacker waits for the victim to make the initial approach and initiate contact. An RSE attack usually requires the attacker to create a persona that would seem attractive to the victim and that would encourage the victim to establish contact. For example, directly calling users and asking them for their passwords on the phone might raise suspicion in some users. In the reverse social engineering version of the same attack, a phone number can be e-mailed to the targets a couple of days in advance by spoofing an e-mail from the system administrator. The e-mail may instruct the users to call this number in case of problems. In this example, any victim who calls the phone number would probably be less suspicious and more willing to share information as she has initiated the first contact.

RSE attacks are especially attractive for online social networks. First, from an attacker's point of view, there is a good potential to reach millions of registered users in this new social setting. Second, RSE has the advantage that it can bypass current behavioral and filter-based detection techniques that aim to prevent wide-spread unsolicited contact. Third, if the victim contacts the attacker, less suspicion is raised, and there is a higher probability that a social engineering attack (e.g., phishing, a financial scam, information theft, etc.) will be successful.

In general, Reverse Social Engineering attacks can be classified based on two main characteristics:

– *Targeted/Un-targeted*: In a targeted attack, the attacker focuses on a particular user. In contrast, in an un-targeted attack, the attacker is solely interested in reaching as many users as possible. Note that in order to perform a targeted attack, the attacker has to know (or acquire) some previous information about the target (e.g., such as her username or e-mail address).
– *Direct/Mediated*: In a direct attack, the baiting action of the attacker is visible to the targeted users. For example, an attacker can post a message on a public forum, or publish some interesting picture on a website. Mediated attacks, in contrast, follow a two-step approach in which the baiting is collected by an intermediate agent that is then responsible for propagating it (often in a different form) to the targeted users.

In the following, we present three different combinations of RSE attacks within the context of online social networks.

Recommendation-Based RSE [Targeted, Mediated] Recommendation systems in social networks propose relationships between users based on background, or "secondary knowledge" on users. This knowledge derives from the interactions between registered users, the friend relationships between them, and other artifacts based on their interaction with the social network. For example, the social networking site might record the fact that a user has visited a certain profile, a page, a picture, and also log the search terms she has entered. Popular social networks (e.g., Facebook) often use this information to make recommendations to users (e.g., "*Visit page X*", "*You might know person Y, click here to become her friends*", etc.).

From an attacker's point of view, a recommendation system is an interesting target. If the attacker is able to influence the recommendation system and make the social network issue targeted recommendations, she may be able to trick victims into contacting her. Figure 1(a) demonstrates the recommendation system-based RSE attack scenario.

Demographic-Based RSE [Un-targeted, Mediated] Demographic-based systems in social networks allow establishing friendships based on the information in a person's profile. Some social networks, especially dating sites (e.g., Badoo), use this technique as the norm for connecting users in the same geographical location, in the same age group, or those who have expressed similar preferences.

Figure 1(b) demonstrates an RSE attack that uses demographic information. In the attack, the attacker simply creates a profile (or a number of profiles) that would have a high probability of appealing to certain users, and then waits for victims to initiate contact.

Visitor Tracking-Based RSE [Targeted, Direct] Visitor tracking is a feature provided by some social networks (e.g., Xing, Friendster) to allow users to track who has visited their online profiles.

Type of Attack	Facebook	Badoo	Friendster
Recommendation-Based	✓✠	-	-
Demographic- Based	✓	✓✠	✓
Visitor Tracking-Based	-	✓	✓✠

Table 1. RSE attacks on three popular social networks. ✓ indicates that the attack is possible; ✠ indicates that we demonstrate and measure the effectiveness of this attack on the particular social network.

The attack in this case involves exploiting the user's curiosity by visiting their profile page. The notification that the page has been visited might raise interest, baiting the user to view the attacker's profile and perhaps take some action. Figure 1(c) outlines this attack method.

3 RSE Attacks in the Real-World

In this section, we present three types of real-world RSE attacks that are possible on three different social network platforms: Facebook, Badoo, and Friendster. In particular, we describe a recommendation-based RSE attack on Facebook, a demographic-based RSE attack on Badoo, and a visitor tracking-based RSE attack on Friendster.

Table 1 shows the social networks that were used in our experiments, and also describes which kind of RSE attacks are possible against them. Note that not all the combinations are possible in practice. For example, Facebook does not provide any information about the users that visit a certain profile, thus making a visitor tracking attack infeasible. In the rest of this section, we describe the different steps that are required to automate the attacks, and the setup of the experiments we performed.

3.1 Ethical and Legal Considerations

Real-world experiments involving social networks may be considered an ethically sensitive area. Clearly, one question that arises is if it is ethically acceptable and justifiable to conduct experiments that involve real users. Similar to the experiments conducted by Jakobsson et al. [11, 12] and our previous work [5], we believe that realistic experiments are the only way to reliably estimate success rates of attacks in the real-world.

Furthermore, during all the experiments we describe in the paper, we took into account the privacy of the users, and the sensitivity of the data that was collected. When the data was analyzed, identifiers (e.g., names) were anonymized, and no manual inspection of the collected data was performed.

Note that all the experiments described in the paper were performed in Europe. Hence, we consulted with the legal department of our institution (comparable to the Institute Review Board (IRB) in the US) and our handling and

privacy precautions were deemed appropriate and consistent with the European legal position.

3.2 Influencing Friend Recommendations

A good example of a real recommendation system is Facebook's friend suggestions. During our tests with Facebook, we observed that Facebook promotes the connection of users by suggesting them friends that they probably know. The system computes these suggestions based on common information, such as mutual friends, schools, companies, and interests. This feature is well-known to many social network users. In fact, whenever a user is logged in, she is regularly notified of persons that she may know.

Previous work [4] has shown that Facebook also uses the e-mail addresses a user has queried to identify a possible friendship connection between two users. The premise is that if users know each other's e-mail addresses, they must be connected in some way. Therefore, if an attacker gains access to the e-mail address of a victim (e.g., a spammer who has a list of e-mails at her disposal), by searching for that address, she can have a fake attacker profile be recommended to the victims. In our experiments, we observed that this technique results in the attacker profile being the most highly recommended profile.

For the first experiment, we used the data collected for over a year in a previous study we performed on Facebook [4]. In the study, we registered a single account that we used to perform a large number of e-mail search queries, using an email list obtained from a dropzone on a machine compromised by attackers. Without our knowledge, our profile was later recommended to all the queried users as a potential friend. As a result, our test account received thousands of messages and friend requests.

3.3 Measuring RSE Effects by Creating Attack Profiles

In the second set of experiments, we created five different attack profiles in three social networks. The profiles were designed with different characteristics to enable us to observe and measure the effects that each characteristic had on the effectiveness of the RSE attacks. That is, we were interested in determining which features would attract the higher number of potential victims using the recommendation-based, demographic-based, and visitor tracking attacks.

The five attack profiles are shown in Table 2. For the profile pictures, we used popular photographs from Wikipedia, licensed under the Creative Commons license. All photos represented an attractive male or female, with the exception of Profile 5 for which we used a synthetic cartoon picture.

Table 3 shows the number of users we targeted in the social networks we tested. For example, in the Facebook experiment, we targeted a total of 250,000 profiles, equally divided between the 5 attack profiles. In the demographic-based attack on Badoo, no action was required on behalf of the attacker. Hence, the number of targeted users is not given (i.e., all registered Badoo users could have found and contacted the attacker profile).

Attribute	Prof. 1	Prof. 2	Prof. 3	Prof. 4	Prof. 5
Age	23	23	23	35	23
Sex	Male	Female	Female	Female	Female
*Location**	N.Y.	N.Y.	Paris	N.Y.	N.Y.
Real Picture	Yes	Yes	Yes	Yes	No

Table 2. Characteristics of the dummy profiles used in the experiments. (* In Badoo, more popular in Europe, we replaced N.Y with London)

Social Network	# of Targets	Total users	Alexia Rank
Badoo	-	73 million	143
Facebook	250,000	500 million	2
Friendster	42,000	8.2 million	643

Table 3. Overview of OSNs as well as number of users targeted.

3.4 Automating the Measurement Process

During our study we developed a number of scripts to automate the three attacks and the measurement process on the different social networks.

Recommendation-Based RSE on Facebook As shown in Figure 1(a), the recommendation-based RSE attack against Facebook consisted of two parts: First, the target user's profile was probed using an e-mail lookup, and second, the attack accounts were automatically monitored for victims who contacted these accounts based on the friendship recommendation made by Facebook.

For the first part, we used the "contact import" functionality provided by Facebook and the API provided by Google Mail's address book to automatically search for users by their e-mail addresses. We broke the total set of users we wished to query into smaller sets, and sent multiple requests to Facebook, as they have limited the number of e-mail addresses that can be queried using a single request (because of recommendations made in previous work [4]).

In the second part of the experiments, we wrote an API that allowed us to interact with Facebook to accept friend requests, fetch user profiles, as well as fetch any private message that may have been sent to the attack profiles.

Note that CAPTCHAs in Facebook were only encountered if we were not careful about rate limiting.

Demographic-Based RSE on Badoo We used Badoo to test the demographic-based RSE attack. Hence, we only had to create the attack profiles and automatically monitor incoming connections. Just like in the recommendation-based RSE attack, we automatically retrieved and collected any message sent to the attacker profiles. Furthermore, as Badoo allows to see which users have visited a profile, we also logged this information.

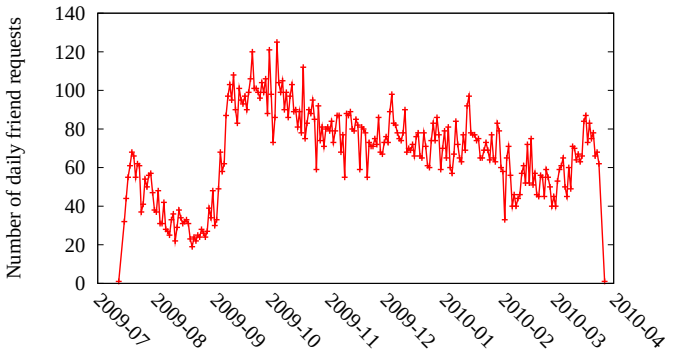


Fig. 2. Daily number of new friend requests in the initial Facebook experiment

Visitor Tracking-Based RSE on Friendster We used Friendster to perform the RSE attack based on visitor tracking. As shown in Figure 1(c), this attack consists of two parts: First, we visit the target user's profile and as a consequence, the system shows to the victim that someone has visited her profile. If the attacker profile is interesting, the victim may choose to contact the attacker. Hence, in a second step, the visits and the incoming messages to the attack profiles were automatically monitored to determine which of the victims came back and initiated contact.

4 Experimental Results

4.1 Recommendation-based RSE Attack

Initial Experiment During the study [4] we conducted, we observed that the test account we were using to query e-mail addresses were receiving a large number of friend requests. The profile used in this attack was similar to Profile 2 described in Table 2.

Figure 2 shows the number of daily friend requests received by the account used in this initial experiment. The graph shows that during the first two months, the account received an average of 45 requests per day, followed by an increase to an average of 75 requests per day for the next 6 months.

The rapid increase in the number of request is the consequence of the cascading effect that commenced when we started accepting the incoming invitations. The fact that the account had a large number of friends built up the "reputation" of our profile. In addition, we started being advertised by Facebook to new people with whom we shared common friends.

Of the over 500,000 e-mails queried by our decoy profile, we were contacted by over 17,000 users (i.e., 3.3% friend connect rate within 9 months and 0.37% friend connect rate per month). Note that our test account reached both the maximum number of active friend connections and the total number of pending friend requests allowed by Facebook.

Controlled, In-Depth Experiments After the success of the initial experiment, we started a number of controlled, in-depth experiments to measure and determine which profile characteristics and social network features affect the success rates of RSE attacks.

To reach our goal, we created five attack profiles on Facebook. For each profile, we randomly selected 50,000 target users and looked up their e-mail addresses (hence, influencing the recommendations made by Facebook). We then measured the number of friend-requests, private messages, and other interaction sent to each attack profile. Figure 3 depicts the result of this experiment. The y-axis represents the cumulative number of friend requests or messages for the period represented by the date on the x-axis.

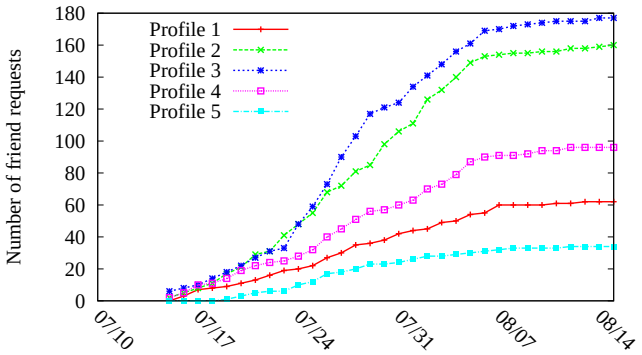
Profiles 2 and 3 were the most successful in terms of the number of friend requests and messages that were received. Both profiles correspond to attractive females who are interested in friendship. Note that there was no correlation with the location of the attack profile (i.e., the location did not influence friend requests). Hence, an initial analysis seems to confirm the general intuition that an attractive female photograph will attract potential victims. In contrast to the other profiles, Profile 5 was the least effective. In this profile, a cartoon character was used as a photograph rather than a real picture. In comparison, Profile 1 performed only slightly better than Profile 5. This profile contained the photograph of an attractive male.

Over the entire month, the most effective profile had a friend connection rate of 0.35% (i.e., in line with the initial experimental profile). The least effective profile instead, had a friend connection rate of only 0.05%.

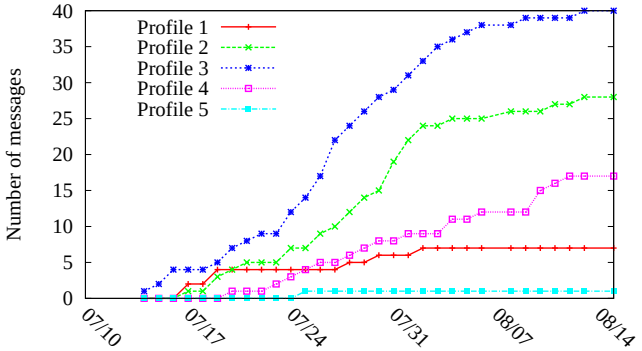
Although friend connection requests and private messages were the most common form of interaction with a decoy profile, we also received a large number of friend suggestions. Friend suggestions are suggestions made by the victim to other users. Such suggestions are important as they imply that a high level of trust has been achieved between the attacker and the victim. Also, note that over 94% of the messages to the attack profiles were sent after the friend connection requests.

By analyzing the demography of the users who contacted our attack profiles, we can identify potential characteristics that make a decoy profile appealing. In particular, we focused on three fields: relationship status, interested in, and age (Figure 4). The y-axis of the figure shows the percentage of friend connection requests that originated from a profile with the respective demographic value (empty values excluded) to the attack profile listed on the x-axis. Young, single users who have expressed interest in "Women" seem to be the easiest victims to attract. In comparison, Profile 1 (the only male profile) received a larger number of friend requests from users who had expressed interest in "Men".

Interestingly, the profile with a cartoon picture was the one to attract the largest number of requests coming from older users (i.e., those who were older than 40). Hence, the experiments show that by carefully tweaking the profile information, it is possible to obtain an higher success rate against a particular group of users.



(a) Friend connect requests sent to each profile



(b) Messages sent to each profile

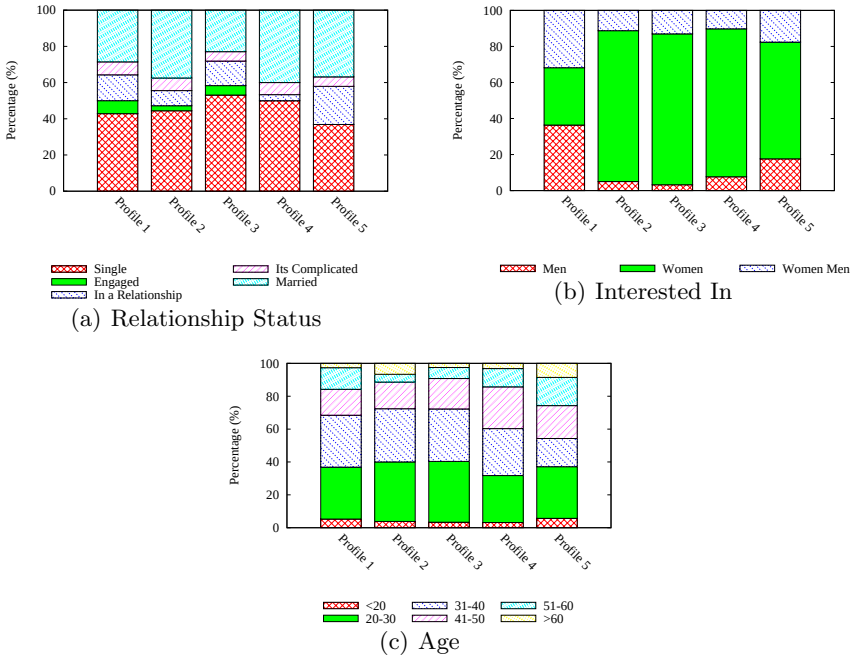
Fig. 3. Cumulative counts of interactions resulting from reverse social engineering on Facebook.

Finally, we analyzed the messages that were sent to the different attack profiles. To protect the privacy of individuals in the study, we first processed the messages and removed user identifiers. After anonymization, we only ran word-based statistical analyses on the message contents. That is, as a pre-processing step, we used Porter's stemming algorithm on the extracted tokens [15], followed by a count of n-grams (where a single gram is a stemmed token).

Around 10% of the messages mentioned the Facebook recommendation, including 3-grams such as "suggest you as" or "suggest I add". The analysis shows that some users used the recommendation made by the social network as a pretext to contact the attack profile.

4.2 Demographic-based Experiment

For our demographic-based RSE attacks, we targeted Badoo, a dating oriented socializing system that allows users to meet new friends in a specific area. A



(a) Relationship Status

(b) Interested In

(c) Age

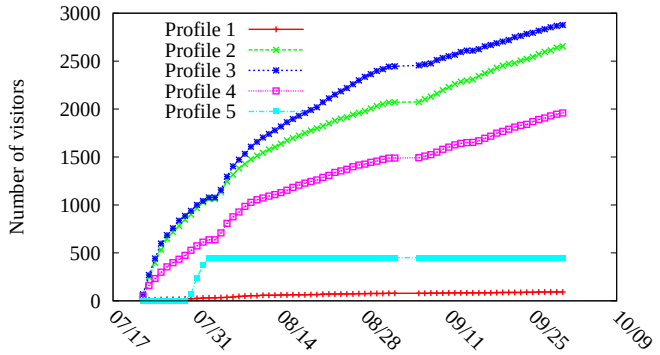
Fig. 4. Demographic breakdown by Relationship Status, Interested In, and Age for Friend Connect requests on Facebook.

registered user can list the people who have visited her profile and exchange messages with other users. Figure 5 shows the cumulative number of visitors and messages received for each attack profile we created in the network.

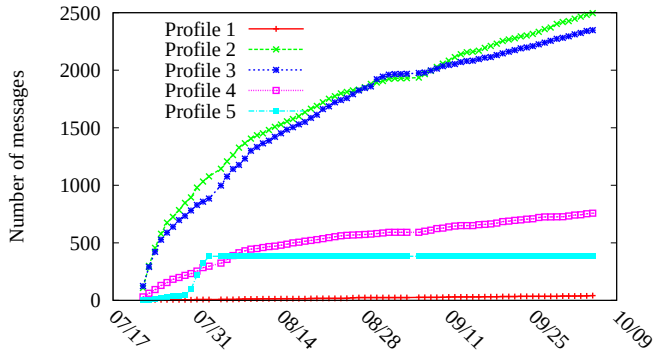
Profiles 2 and 3 were again the most popular, and attracted the most visitors (over 2500 each). These profiles also received the largest number of messages (i.e., more than 2500 each). Because Profile 5 was not using a photograph of a person, it was removed by Badoo from the demographic search after it was visited by 451 users and it received 383 messages. Once again, Profile 1, the attack profile of a male user, received the fewest visits and friend requests.

Another measure of how successful an attack profile was is the percentage of users who decided to send a message after visiting a profile. These figures are over 50% for the two attractive female profiles (Profile 2 and 3), and 44% on average for all attack profiles.

We took a closer look at the demography of the users who contacted us. In the case of Badoo, sending a message is the most concrete form of interest, and one that can easily be exploited (e.g., [5]). Figure 6 shows a demographic breakdown by relationship status, what users were interested in, and age. Similar to Figure 4, the y-axis shows the percentage of users who sent messages that originated from a profile with the respective demographic value.



(a) Visitors to each profile



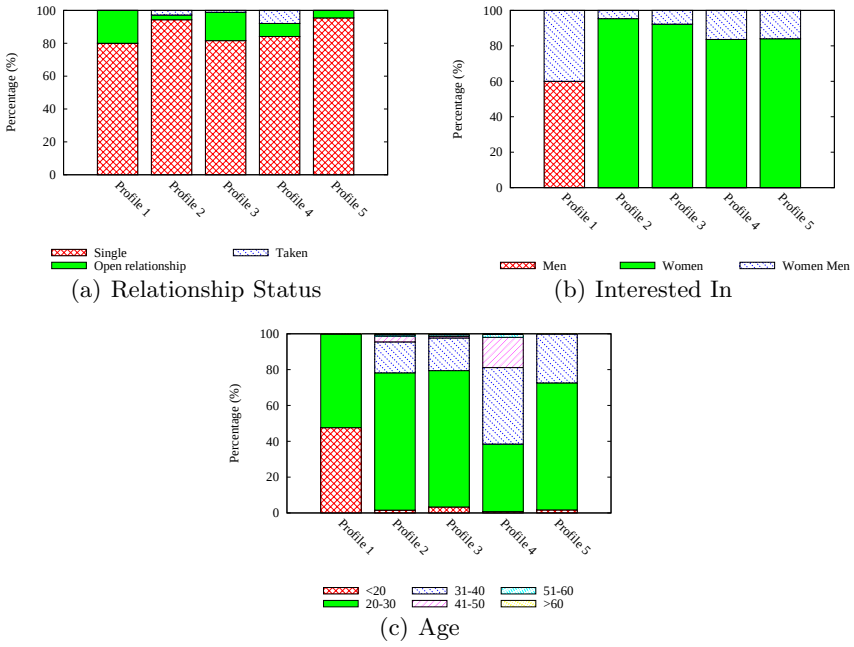
(b) Messages sent to each profile

Fig. 5. Cumulative counts of interactions resulting from reverse social engineering on Badoo.

Note that Badoo is a site that is geared towards dating. Most of the users who initiate contact express that they are either single, or in an "open relationship". In general, the attack profiles only attracted users of the opposite gender. The age demographic shows that most of the victims belong to the same age group that the attack profile belongs to. In comparison, there was no correlation of age for contact requests on Facebook.

Another important difference with respect to Facebook was that the location was significant in Badoo. In fact, almost all the messages were sent by people living in the same country as the attack profile.

Finally, the 3-grams analysis for the messages received on Badoo showed that the most popular term was "how are you" occurring over 700 times. Other popular lines included "get to know" and "would you like", "you like" ... "chat" or "meet".



(a) Relationship Status (b) Interested In

(c) Age

Fig. 6. Demographic breakdown by Relationship Status, Interested In, and Age for messages on Badoo.

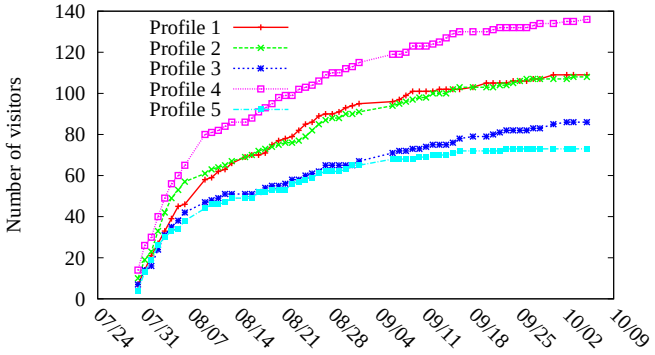
4.3 Visitor Tracking Experiment

In the visitor tracking RSE attack, we used each of the five attack profiles to visit 8,400 different user profiles in Friendster. As we have already previously described, on Friendster a user can check which other users have visited her profile.

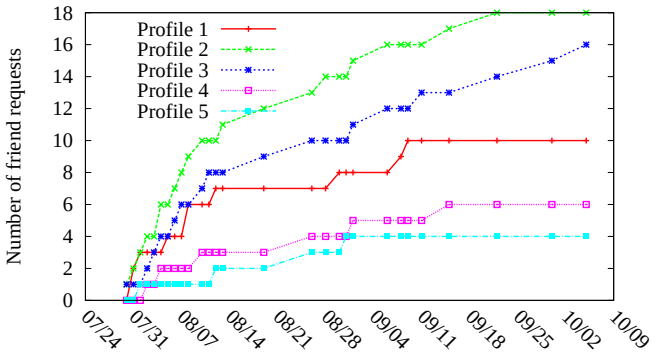
In our experiment, we tracked which victims visited our attack profiles, and then counted the number of users who sent us a friend request. The results of this experiment are shown in Figure 7 (the sub-figure 7(a) and 7(b) represent the number of visitors and number of friend requests sent to the attack profiles).

The number of users who were curious about our visit, and visited us back was consistent with the results of the experiments we conducted on other social networks (i.e., between 0.25 and 1.2% per month). However, only a few users later sent a friend request or a message.

The demographic breakdown for Friendster is presented in Figure 4.3. The statistical distributions are similar to the ones obtained in the Facebook experiment, proving the difference in terms of characteristics between friend-oriented and dating-oriented social networks.



(a) Visitors to each profile



(b) Friend requests sent to each profile

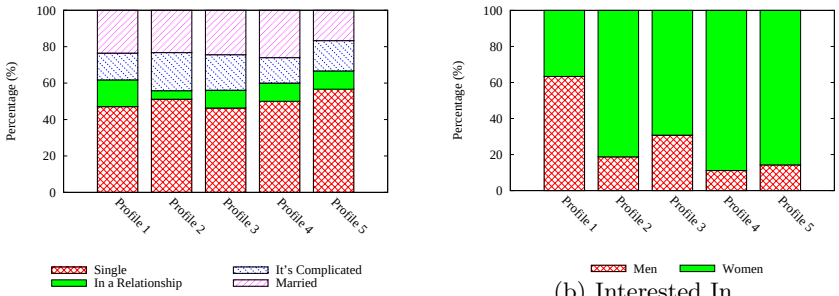
Fig. 7. Cumulative counts of interactions resulting from reverse social engineering on Friendster.

5 Discussion and Lessons Learned

In this section, based on the results of the empirical experiments, we distill some insights about the way RSE attacks work in social networks. We can summarize our findings in two main points: The importance of having the right profile, and the importance of providing a pretext to the victims.

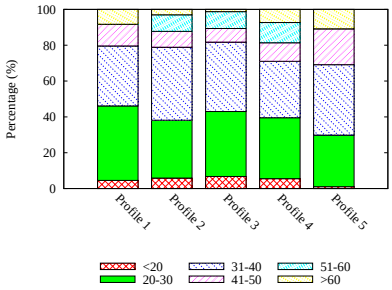
The first, straightforward, factor we were able to measure is the impact of the profile characteristics on the overall effectiveness of an attack. The experiments confirm the folk wisdom that using an attractive female photograph is a good choice to attract victims. The success rate of the most successful female profile, in terms of both friend requests and number of received messages, is between 2 and 40 times higher than the worse performing profiles (i.e., the male profile and the profile without a photograph).

Note that if the objective of the attack is not simply to reach the highest number of users, but to target a specific person, or group, the success rate of the



(a) Relationship Status

(b) Interested In



(c) Age

attack can be improved by carefully tuning the profile characteristics. For example, our experiments show that age and location information are decisive in dating sites, while this information is not as critical in more general, friend-oriented, social networks. Also, the results suggest that gender information is always very important. Hence, a successful reverse social engineering attack should use the opposite sex of the victims in the decoy profile.

The experiments show that the impact of the profile picture is quite uniform in different social networks. For example, we observe that young users are generally more intrigued by attractive photographs, while decoy profiles (e.g., Profile 5) that do not contain the photograph of a real person tend to attract more senior users.

Obviously, even though having a catchy, interesting profile is important, our research shows that there is a second, even more important factor that contributes to the success of the attack: the pretext. Our experiments indicate that users need an incentive and a good reason to engage in interaction with a person that they do not know. In other words, users need a good excuse to "break the ice" and motivate the first approach. The differences between the success rates of the attacks on Facebook and Friendster suggest that an incentive or a pretext is critical for reverse social engineering attacks to work in practice.

The analysis of the messages received on Facebook support the hypothesis that a recommendation system gives a reason to users to initiate contact. That is, a number of users referenced the Facebook recommendation as a motivation

for their friend request. In contrast, on Friendster, even though the percentage of users that browsed our decoy profiles was consistent with the other social network experiments, very few people moved to the next step and sent a contact message. The reason is, in our opinion, that the visitor tracking attack failed to provide a good pretext to the victims.

Note that the demographic experiment on Badoo was also very effective. The reason for this success is that Badoo greatly relies on the demographic search functionality to allow users to find possible contacts. In the case of a dating site, the pretext for establishing contact was the fact itself of living in a close location, or being in the same age group of the victim.

Our experiments demonstrate that reverse social engineering attacks on social networks are feasible if they are properly designed and executed. However, contrary to the common folk wisdom, only having an account with an attractive photograph may not be enough to recruit a high number of unsuspecting victims. Rather, the attacker needs to combine an attractive profile with a pretext and incentive for the victim to establish contact. Recommendation systems such as Facebook's friend suggestions are effective tools for creating such an incentive. Also, we see that profile attributes such as location and age may be the required incentives on dating networks such as Badoo.

6 RSE Countermeasures in OSN

Clearly, features that allow social network users to easily make new acquaintances are useful in practice. However, our paper demonstrates that such systems may also be abused to trick users on behalf of attackers. In this section, we list three countermeasures that would increase the difficulty of launching RSE attacks in online social networks.

First, while friend recommendation features are useful, our experiments show that they may pose a risk to users if the attackers are able to somehow influence the recommendation system. Hence, it is important for social network providers to show a potential connection between two users only if there is a strong connection between them. For example, in the case of Facebook, as our experiments show, a simple e-mail lookup does not necessarily indicate that the users know each other. Thus, one could check other information, such as the fact that the users already have some friends in common.

Second, we believe that it is important to closely monitor friendships that have been established in social networks. Benign user accounts will typically send and receive friend requests in both directions. That is, a user may be contacted by people she knows, but she will also actively search and add friends on the network. However, in contrast, a honeypot RSE account (as we describe in this paper) only receives friend requests from other users. Thus, it may be possible to identify such accounts automatically.

Third, we believe that CAPTCHA usage also needs to be extended to incoming friend requests. Today, because of the active threats of spamming and social

engineering, social network providers may display CAPTCHAs when friend requests are sent to other users. However, no such precautions are taken for messages and friend requests that are received. By requiring to solve a CAPTCHA challenge before being able to accept suspicious incoming friend requests, we believe that RSE attacks would become more difficult. While CAPTCHAs are not the silver bullet in preventing and stopping malicious activity on social networks (e.g., as show in [1, 5]), they do raise the difficulty bar for the attackers.

7 Related Work

Social engineering attacks are well-known in practice as well as in literature (e.g., [14, 3, 17, 8, 16]). Social engineering targets human weaknesses instead of vulnerabilities in technical systems. Automated Social Engineering (ASE) is the process of automatically executing social engineering attacks. For example, spamming and phishing can be seen as a very simple form of social engineering (i.e., making users click on links).

A general problem on social networks is that it is difficult for users to judge if a friend request is trustworthy or not. Thus, users are often quick in accepting invitations from people they do not know. For example, an experiment conducted by Sophos in 2007 showed that 41% of Facebook users acknowledged a friend request from a random person [1]. More cautions users can be tricked by requests from adversaries that impersonate friends [5]. Unfortunately, once a connection is established, the attacker typically has full access to all information on the victim's profile. Moreover, users who receive messages from alleged friends are much more likely to act upon such message, for example, by clicking on links. A similar result was reported by Jagatic et al. [10]. The authors found that phishing attempts are more likely to succeed if the attacker uses stolen information from victims' friends in social networks to craft their phishing e-mails.

In contrast to active social engineering that requires the attacker to establish contact with the victim, in a reverse social engineering attack, it is the victim that contacts the attacker. We are not aware of any previous reports or studies on reverse social engineering attacks in online social networks. The results of this paper demonstrate that automated reverse social engineering is a realistic threat, and that it is feasible in practice.

The most well-known attack to compromise the trust relationship in a social network that employs a reputation system is the *sybil attack* [6]. In this attack, the attacker creates multiple fake identities and use them to gain a disproportionately large influence on the reputation system. Note that the findings in this paper have implications for research that aims to defend social networks against sybil attacks (e.g., SybilGuard [18], SybilLimit [19]). SybilGuard and SybilLimit assume that real-world social networks are fast mixing [7] and this insight is used to distinguish the sybil nodes from normal nodes. Fast mixing means that subsets of honest nodes have good connectivity to the rest of the social network. Both SybilGuard and SybilLimit are good solutions for detecting Sybil nodes. However, the attacks we present in this paper result in legitimate friend-

ship connections and, therefore, would not be detected by current sybil-detection approaches.

8 Conclusion

Hundreds of millions of users are registered to social networking sites and regularly use them features to stay in touch with friends, communicate, do online commerce, and share multimedia artifacts with other users.

To be able to make suggestions and to promote friendships, social networking sites often mine the data that has been collected about the registered users. For example, the fact that a user looks up an e-mail address might be assumed to indicate that the user knows the person who owns that e-mail account. Unfortunately, such assumptions can also be abused by attackers to influence recommendations, or to increase the chance that the victim's interest is intrigued by a fake honey-account.

Although social engineering attacks in social networks have been well-studied to date, *reverse social engineering* (RSE) attacks have not received any attention.

This paper presents the first user study on how attackers can abuse some of the features provided by online social networks with the aim of launching automated reverse social engineering attacks. We present and study the effectiveness and feasibility of three novel attacks: Recommendation-based, visitor tracking-based, and demographic-based reverse social engineering.

Our results show that RSE attacks are a feasible threat in real-life, and that attackers may be able to attract a large numbers of legitimate users *without* actively sending any friend request. The experiments we have conducted demonstrate that suggestions and friend-finding features (e.g., demographic-based searches) made by social networking sites may provide an incentive for the victims to contact a user if the right setting is created (e.g., an attractive photograph, an attack profile with similar interests, etc.).

We hope that this paper will raise awareness about the real-world threat of reverse social engineering in social networks and will encourage social network providers to adopt some countermeasures.

Acknowledgments. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n^o 257007. This research has been partially funded by National Science Foundation by IUCRC, CyberTrust, CISE/CRI, and NetSE programs, National Center for Research Resources, and gifts, grants, or contracts from Wipro Technologies, Fujitsu Labs, Amazon Web Services in Education program, and Georgia Tech Foundation through the John P. Imlay, Jr. Chair endowment. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding agencies and companies mentioned above.

References

[1] Sophos Facebook ID Probe. http://www.sophos.com/pressoffice/news/articles/2007/08/facebook.html, 2008.

[2] Facebook Statistics. http://www.facebook.com/press/info.php?statistics, 2010.

[3] Sophos Security Threat 2010. http://www.sophos.com/sophos/docs/eng/papers/sophos-security-threat-report-jan-2010-wpna.pdf, 2010.

[4] Balduzzi, M., Platzer, C., Holz, T., Kirda, E., Balzarotti, D., and Kruegel, C. Abusing Social Networks for Automated User Profiling. In *Recent Advances in Intrusion Detection* (2010), Springer, pp. 422–441.

[5] Bilge, L., Strufe, T., Balzarotti, D., and Kirda, E. All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks. In *18th International Conference on World Wide Web (WWW)* (2009).

[6] Douceur, J. R. The Sybil Attack. In *Electronic Proceedings for the 1st International Workshop on Peer-to-Peer Systems (IPTPS '02)* (March 2002).

[7] Flaxman, A. Expansion and lack thereof in randomly perturbed graphs. *Internet Mathematics 4*, 2 (2007), 131–147.

[8] Irani, D., Webb, S., Giffin, J., and Pu, C. Evolutionary study of phishing. In *eCrime Researchers Summit, 2008* (2008), IEEE, pp. 1–10.

[9] Irani, D., Webb, S., Pu, C., and Li, K. Study of Trend-Stuffing on Twitter through Text Classification. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)* (2010).

[10] Jagatic, T. N., Johnson, N. A., Jakobsson, M., and Menczer, F. Social phishing. *Commun. ACM 50*, 10 (2007), 94–100.

[11] Jakobsson, M., Finn, P., and Johnson, N. Why and How to Perform Fraud Experiments. *Security & Privacy, IEEE 6*, 2 (March-April 2008), 66–68.

[12] Jakobsson, M., and Ratkiewicz, J. Designing ethical phishing experiments: a study of (ROT13) rOnl query features. In *15th International Conference on World Wide Web (WWW)* (2006).

[13] Lauinger, T., Pankakoski, V., Balzarotti, D., and Kirda, E. Honeybot, your man in the middle for automated social engineering. In *LEET'10, 3rd USENIX Workshop on Large-Scale Exploits and Emergent Threats, San Jose* (2010).

[14] Mitnick, K., Simon, W. L., and Wozniak, S. *The Art of Deception: Controlling the Human Element of Security*. Wiley, 2002.

[15] Porter, M. An algorithm for suffix stripping. *Program 14*, 3 (1980), 130–137.

[16] Stringhini, G., Kruegel, C., and Vigna, G. Detecting Spammers on Social Networks. In *Annual Computer Security Applications Conference (ACSAC)* (2010).

[17] Webb, S., Caverlee, J., and Pu, C. Social Honeypots: Making Friends with a Spammer Near You. In *Conference on Email and Anti-Spam (CEAS)* (2008).

[18] Yu, H., Kaminsky, M., Gibbons, P., and Flaxman, A. Sybilguard: defending against sybil attacks via social networks. In *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications* (2006), ACM, pp. 267–278.

[19] Yu, H., Kaminsky, M., Gibbons, P. B., and Flaxman, A. SybilLimit: A Near-Optimal Social Network Defense against Sybil Attacks. In *IEEE Symposium on Security and Privacy* (2008).