Archives of Business Research – Vol.7, No.3
Publication Date: Mar. 25, 2019
DOI: 10.14738/abr.73.6265.
Algarni, A. A., Almakrami, H., & Alarifi, A. H. (2019). Toward Evaluating Trustworthiness of Social Networking Site Users: Reputation-Based Method. *Archives of Business Research,* 7*(3),* 27-41.

Toward Evaluating Trustworthiness of Social Networking Site Users: Reputation-Based Method

Abdullah Ayed Algarni
Information Technology Division,
Institute of Public Administration, Riyadh, Saudi Arabia.

Hashem Almakrami
Information Technology Division,
Institute of Public Administration, Riyadh, Saudi Arabia.

Abdulrahman H. Alarifi
Information Technology Division,
Institute of Public Administration, Riyadh, Saudi Arabia.

ABSTRACT

As social networking sites (SNSs) have risen in popularity, attackers have been using social engineering traps and tactics to trick SNS users into obeying them, accepting threats, and falling victim to various crimes and attacks, such as phishing, sexual abuse, financial abuse, identity theft, impersonation, physical crime, and many other forms. Recent research on SNS security shows that most of the attackers rely mainly on fake identities. However, one of the key challenges that has faced researchers recently is how to distinguish between legitimate users and attackers. In this paper, we propose a simple yet effective method of evaluating the trustworthiness of an SNS user. The proposed method relies on a user's reputation, which can be evaluated from the user's friendship history. As such, this method contributes to reducing the risks associated with the lack of identity authentication in SNSs, as well as the failure to filter fake profiles when receiving friendship invitations, looking for people on search engines, and dealing with spam messages.

Key words: Social engineering, deception, source credibility, phishing, social networking sites.

INTRODUCTION

Security threats in information systems generally occur through the vulnerabilities of technologies or of people. People are considered the weakest links in security (Nohlberg, 2009; West, Mayhorn, Hardee, & Mendel, 2009). *Social engineering* is the art of deceiving or tricking people to obtain information from them or to persuade them to perform an action that benefits the attacker in some way (Hadnagy, 2010; Thornburgh, 2004; Workman, 2007). Many organizations recognize the importance of predicting and controlling social engineering, but many fail to reach this goal (Brody, 2012).

Recently, fraudulent and deceptive people have been deploying social engineering traps and tactics by using social networking sites (SNSs) to trick victims into obeying them, accepting threats, and falling victim to various crimes and attacks, such as phishing, sexual abuse, financial abuse, identity theft, impersonation, physical crime, and many other forms. Several researchers have investigated and highlighted the risks associated with social engineering in SNSs (e.g., Algarni, Xu, Chan, & Tian, 2013a, 2013b; Braun & Esswein, 2013; Chitrey, Singh, & Singh, 2012; Dimensional-Research, 2011; Jagatic, Johnson, Jakobsson, & Menczer, 2007; Nagy & Pecho, 2009). These studies have suggested that SNSs are currently the most common

sources of social engineering threats. The simple trick of offering free cell phone minutes accounted for the largest number of attacks on Facebook users in 2013, increasing from 56% in 2012 to 81% in 2013 (Laura Mazzuca, 2014). Recent research on SNS security shows that most social engineering threats, such as spamming, identity cloning, and social bots, rely mainly on fake identities (Fire, Goldschmidt, & Elovici, 2014). This fact explains why an estimated 83 million (8.7% of all accounts) Facebook accounts may be fake (Couper, 2013).

The risk of social engineering attacks in SNSs is associated with how difficult it is for users to make accurate judgments regarding identity trustworthiness in the virtual environment. Despite the previous mentioned studies that attempt to provide solutions that can help SNS users or providers to identify fake profiles, such solutions seem complicated and difficult to apply in the actual SNSs. Friendship requests are still received in most SNSs, such as Facebook, without any indication of the senders' trustworthiness. No effective filters or trustworthiness evaluators are used in the majority of SNSs, making SNSs the perfect breeding grounds for malicious users and attackers and making users susceptible to many social engineering threats. In this paper, we propose a simple yet effective method of evaluating the trustworthiness of SNS users. As such, we hope that this method will help reduce the risks associated with the lack of identity authentication in SNSs and the failure to filter fake profiles when receiving friendship invitations, browsing search engines, and dealing with spam messages.

LITERATURE REVIEW

The topic of social engineering in SNSs has attracted many researchers in recent years. Several studies have investigated and highlighted the risks associated with social engineering in SNSs (e.g., Boorman et al., 2014; Chitrey et al., 2012; Dimensional-Research, 2011; Fire et al., 2014; Hogben, 2007; Jagatic et al., 2007; Krombholz, Hobel, Huber, & Weippl, 2014; Nagy & Pecho, 2009; Shariff & Zhang, 2014). These studies have suggested that SNSs are the most common sources of social engineering attacks.

Many researchers have tried to develop solutions to overcome the problems associated with social engineering in SNSs. One research track that has tried to provide solutions to the social engineering issue involves spam detection. Spam detection solutions attempt to identify if a message (tweet, comment, or post) is legitimate or spam. These solutions rely mainly on measuring the similarity among messages. Spam messages are usually similar in their specifications. The relationship between spamming and social engineering is that spam messages can contain social engineering tricks. Gao et al. (2010) studied spamming messages and found that 70% of all malicious wall posts advertised phishing sites. Social engineering can also be used to trick users to give spammers hidden permission to post on their walls or send spam messages to their friends. The major solutions proposed in the literature are based on similarity features. For example, Stringhini, Kruegel, and Vigna (2010b) studied the characteristics of spam messages to be able to detect them automatically. They found that spam messages usually took the form of advertisements and contained URL links to particular websites. They proposed a technique to detect spammers in social networks. While their proposed technique can work on greedy bots that send spam with each message, a low-traffic spamming campaign would not be easy to detect.

Rahman, Huang, Madhyastha, and Faloutsos (2012) also used the similarity features to detect spam messages. They calculated the similarity score that summed the value for all similar messages, with the same URL links, and computed the standard deviation for all the posts. This technique seems to offer a better solution than that of Stringhini et al. (2010b), as the former carefully checks if the message is spam. Some researchers (Benevenuto, Magno, Rodrigues, &

Almeida, 2010; Castillo, Mendoza, & Poblete, 2011; Huber, Kowalski, Nohlberg, & Tjoa, 2009; McCord & Chuah, 2011; Stringhini, Kruegel, & Vigna, 2010a; Wang, 2010; Yardi, Romero, & Schoenebeck, 2009) investigated spam detection on Twitter by using a content-based approach and achieved different accuracy results. Their approaches were mainly based on similarity and the characteristics of the URL links. The major limitation of these automatic methods is that they do not work effectively in detecting low-traffic spamming campaigns.

Another research track involves the detection of bot-operated accounts. Its solutions aim to detect if a profile is operated by a human or a computer (bot). Bots are "automatic or semi-automatic computer programs that mimic humans and/or human behaviour in online social networks" (Wagner, Mitter, Körner, & Strohmaier, 2012, p. 41). Detecting this type of account or profile is mainly based on the types of posts and messages that are created by an account. If the account posts or sends messages that can be classified as spam, the algorithm suggests that the operator is a computer. However, if the profile has normal communication with friends and various social transactions, the algorithm suggests that the operator is human. Several researchers (e.g., Castillo et al., 2011; Chu, Gianvecchio, Wang, & Jajodia, 2012; Huber et al., 2009; McCord & Chuah, 2011; Stringhini et al., 2010a; Wang, 2010) used content-based techniques, similar to those described above in relation to spam detection, to detect if the account operation was automated or performed by a human being.

One of the important studies in this regard was conducted by Yardi et al. (2009), who examined the differences between fake and legitimate Twitter users, mainly based on content-based techniques. Chu et al. (2012) proposed an improved classification model that categorized legitimate users, automated users (bots), and a combination of both on Twitter. While their proposed classification model shows high accuracy when using all methods that operate the algorithm, it seems costly, with high computational complexity. The reason is that it includes many factors such as the URL ratio, mention ratio, registration date, link safety, hashtag ratio, follower-to-friend ratio, and account verification. Moreover, their classification model can work only for automated programs (i.e., bots) and cannot be generalized to any fake profile.

Fire, Katz, and Elovici (2012) proposed a novel algorithm for detecting malicious profiles in SNSs. Their algorithm uses a combination of graph theory algorithms and machine learning to detect malicious profiles that can be classified as spammers. Their algorithm has the advantage of being evaluated on several SNSs and found effective in detecting spammers' profiles. Thomas, McCoy, Grier, Kolcz, and Paxson (2013) investigated the market for fraudulent Twitter accounts (profiles) to monitor the fraud perpetrated by 27 merchants over a 10-month period. They were able to monitor around 120,000 fraudulent accounts. Based on their exploration, they developed a classifier to retroactively detect the fraudulent accounts sold via these merchants. Their work is unique in the area of SNS spamming. However, their proposed classifier technique is still mostly helpful only in relation to automatically generated accounts. Their classifier was developed and tested on Twitter, and whether or not their classifier algorithm works for any SNS or only for Twitter requires further investigation.

Abbasi and Liu (2013) designed an algorithm called CredRank to measure information source credibility on social media. This algorithm relies on examining the similarities in behaviors (not messages) of SNS users. They built their algorithm based on two assumptions: 1) A non-credible user creates a large number of accounts and uses these to spread messages or words. 2) A non-credible user votes, regardless of content, for other users in its group. The first assumption had been supported by several studies that investigated spam messages' behaviors. However, there seems to be a lack of theories that support the second assumption.

Nevertheless, the goal of Abbasi and Liu's (2013) algorithm is to identify the similarities among users and cluster them. The clusters are then weighted to show the credibility value of their members, which helps detect any coordinated behavior, such as a fake profile used for spamming or a Sybil profile used to manipulate the voting or rating system. While the idea behind this proposed algorithm is novel, it is limited to predicting the credibility of the information in terms of spamming, based on classifying the profile (i.e., whether the profile is a bot or a human), but it cannot be generalized to the credibility of the source. Source credibility is a complex and multidimensional concept (Eisend, 2006), and limiting it to classifying the profile as a computer or a human is an approach that lacks evidence.

Additionally, Conti, Poovendran, and Secchiero (2012) introduced a new model that attempted to mitigate fake account (fake profile) attacks. Their model depends on the temporal evolution that characterizes real SNS user accounts, whereby the data can be collected and utilized to identify a set of features in the dynamic mode of SNSs. These features can be utilized to evaluate a particular profile being tested and to detect if there is any major deviation from expected behavior. Meligy, Ibrahim, and Torky (2015) proposed a theoretical framework that mainly relied on a novel topology named the "trusted social graph." Their approach aimed to visually show the trusted instances of social interactions among users and to detect the strange instances of communications that are more likely to be performed by cloned profiles. Furthermore, several researchers (Al Zamal, Liu, & Ruths, 2012; Burger, Henderson, Kim, & Zarrella, 2011; Liu & Ruths, 2013); Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011; Rao et al., 2011) proposed solutions that tried to classify the identity (usually the gender) of SNS profile owners. Their solutions attempted to detect the false information provided in users' profiles. Despite the importance and creativity of these works, detecting deception involving the identity (e.g., the gender) of SNSs is still challenging. To date, no reliable solution has been devised for detecting social engineering of this kind.

ASSUMPTIONS

The proposed method relies mainly on the following three assumptions:

Assumption 1. Most of the SNSs' legitimate users tend to accept friendship invitations/requests from real (offline) friends.

Assumption 2. Most of the SNSs' legitimate users tend to reject friendship invitations/requests from strangers.

Assumption 3. Most of the SNSs' legitimate users tend to send friendship invitations/requests to real (offline) known friends.

VALIDATION OF ASSUMPTIONS

Validation Objective and Procedures

This step aims to validate the three identified assumptions, based on users' opinions. To obtain better results, we conducted a qualitative questionnaire-based online survey to collect and understand people's experiences in accepting or rejecting friendship invitations. The qualitative questionnaire-based survey is a technique whereby the researcher gains a deep understanding of human behaviors, as well as the different reasons that govern these behaviors (Denzin & Lincoln, 2005). The qualitative method involves investigations into the how and the why of decision making rather than just focusing on when, where, and what questions.

To fulfill our research objective, we administered a questionnaire and collected the participants' insightful opinions. To recruit more participants, we made the survey concise and the participation anonymous. Moreover, to avoid fabricated stories or bias, we made the

participation totally voluntary. In the beginning of the survey, the respondents were given information in a short paragraph to illustrate what we meant by SNSs and friendship invitations, as well as to explain the purpose of the study. Demographic variables were provided in a drop-down list to choose from, and the following questions were asked:

1. What do you use SNSs for?
2. What kinds of friendship invitations do you usually accept on SNSs? What criteria would you use in your decision?
3. What kinds of friendship invitations do you usually reject on SNSs? What criteria would you use in your decision?
4. What kinds of SNS users are you the most eager to connect to (send friendship invitations to them)?

Validation Sampling and Analysis

Around 800 people were approached, and 477 responses were collected. The recruited sample included both genders, a wide age range (18–60 years old), and a variety of educational levels (secondary school and bachelor's, master's, and PhD degrees). After critical screening, five respondents were not relevant as they indicated that they used SNSs for illegitimate purposes, such as fraud or hacking; therefore, they were discarded. Thematic analysis—a technique whereby the researcher identifies themes or patterns in the data that are thought to reflect the participants' experiences—was then used to analyze the data (Braun & Clarke, 2006). The data obtained from the questionnaire-based survey and the thematic analysis were handled by using the manual method of color coding and noting to identify and group ideas or nuances that appeared to be connected to the subject under investigation (Roberts & Taylor, 2002).

Validation Findings

For the first question of the survey, the majority (472 out of 477) of the participants indicated that they used SNSs for communication with friends and family members, reading news, playing games, and other legitimate purposes. For the second question, more than 90% of the participants answered that they tended to accept friendship invitations/requests from real (offline) friends. The data collected regarding the third question showed that around 80% of the respondents usually rejected friendship invitations from strangers. Similarly, around 85% of the participants noted that they used the same criteria when deciding to send friendship invitations. They were eager to connect to those people whom they knew offline (in real life).

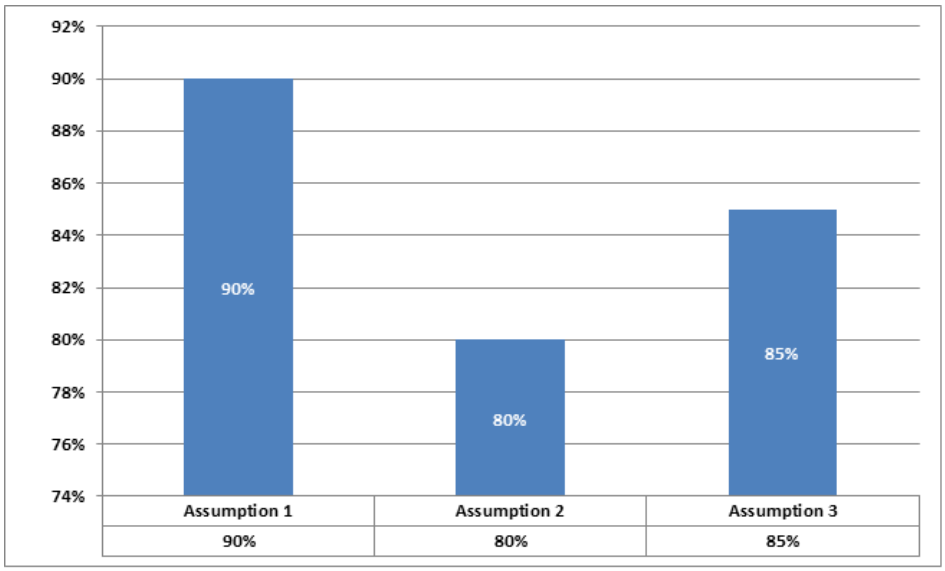

Figure 1. Assumption validation

Figure 1 shows the percentage of the participants' responses regarding each assumption. It can be concluded that all of the three assumptions have been validated. Furthermore, we can extend these assumptions as follows:

Assumption 4. Receiving friendship requests from other users means that these senders are more likely to know their intended recipients. Therefore, receiving a friendship request from a highly trustworthy sender gives the recipient a higher trustworthiness rate than receiving a friendship request from a sender with a lower trustworthiness rate.

Assumption 5. Rejecting friendship requests that a user sends to others means that those users who reject the friendship requests more likely do not know the sender. Therefore, rejecting the friendship request sent by a user decreases the trustworthiness of that user.

PROPOSED METHOD

The *reputation-based detection method* (RDM) aims to estimate a given user's degree of trustworthiness in an SNS based on the user's friendship history. Tracking the user's friendship invitations (acceptance, rejection, sending, and receiving) could provide a simple yet reliable method of evaluating his or her trustworthiness and thus detecting a suspicious account, which could be a fraud, a spammer, a deceptive user, or any fake profile.

The RDM method estimates each user's degree of trustworthiness from 0 to 10, based on the user's friendship invitation history, as follows:

- The lowest rate is 0, which indicates that the user is extremely untrustworthy.
- The highest rate is 10, which indicates that the user is highly trustworthy.
- A rate of 5 means that the user has no indication of being trustworthy or not.

In other words, as a given user's trustworthiness rate increases, this user's potential risk decreases. Conversely, as a given user's trustworthiness rate declines, this user's potential risk increases.

HOW DOES THE RDM METHOD WORK?

To provide a better understanding of the RDM method, the following scenario is presented:

1. We represent each user by a node with a given name. In the example shown in Figure 2, the nodes are named X, A, B, C, D, and E.
2. Each node is connected to one or more nodes through a link, which represents a friendship between a given user and one or more other users. In the example shown in Figure 2, node X is connected to nodes A, B, and C. This means that the user who is represented by node X is a friend of the users who are represented by nodes A, B, and C.
3. Each arrow in Figure 2 represents the direction of the friendship request/invitation. The arrow originates from the user who requests (sends) the friendship (invitation). In the example shown in Figure 2, friendship requests have been sent by the user who is represented by node X to the users who are represented by nodes A and C, respectively. Similarly, a friendship request has been sent by the user who is represented by node B to the user who is represented by node X.
4. Each dashed line represents a rejected friendship request. The dashed line between nodes X and D indicates that the user who is represented by node D has rejected the friendship request sent by the user who is represented by node X. The dashed line between nodes X and E illustrates a similar case, with the former's rejection of the latter.

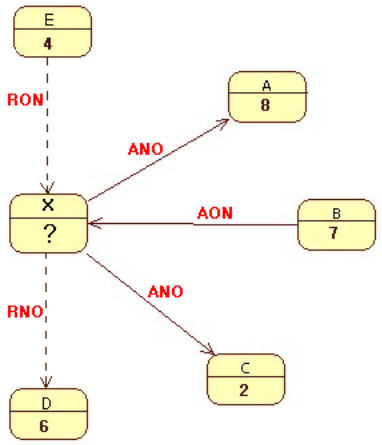

Figure 2. RDM method—first scenario

5. Figure 2 shows four types of friendship requests, as follows:
   a. accepted node to others (ANO) (e.g., between nodes X and A),
   b. accepted others to node (AON) (e.g., between nodes B and X),
   c. rejected others to node (RON) (e.g., between nodes E and X), and
   d. rejected node to others (RNO) (e.g., between nodes X and D).

Considering the assumptions that have been explained in the Assumptions section, the RDM method estimates a given user's trustworthiness rate, based on the first three types of friendship requests, through the following equation:

The trustworthiness rate of a given node = **[5 * ((total weight of ANO) / ((total weight of ANO) + (total weight of RNO))] + [5 * ((total weight of AON) / ((number of AON * 10))].**

As shown in Figure 2, suppose that the nodes' trustworthiness rates are A = 8, B = 7, C = 2, D = 6, and E = 4, then the RDM method estimates node X's trustworthiness rate as follows:

Trustworthiness rate of node X = [5 * (10 / (10 + 6))] + [5 * (7 / (1 * 10))] = 6.625

ADDITIONAL SCENARIOS

In this section, we consider other possible scenarios to check the efficiency of the proposed method.

1. As shown in Figure 3, the user represented by node X has sent five friendship requests to the users represented by nodes A, B, C, D, and E. All of these requests have been accepted.

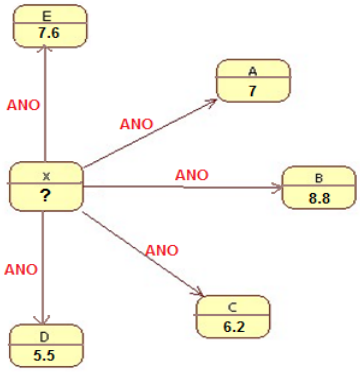

Figure 3. RDM method—second scenario

Based on the RDM method, node X's trustworthiness rate = 5. Node X's trustworthiness rate of 5 is reasonable since there is no indication of its being trustworthy (no user has sent X a friendship request) and none for being untrustworthy (no user has rejected any friendship request sent by X).

2. As shown in Figure 4, the user represented by node X has sent five friendship requests to the users represented by nodes A, B, C, D, and E. Three of these requests have been accepted, and two have been rejected.

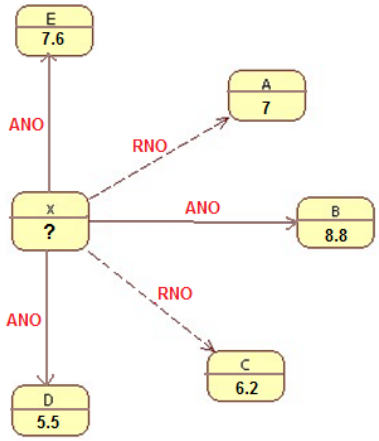

Figure 4. RDM method—third scenario

Based on the RDM method, node X's trustworthiness rate = 3.12. Node X's trustworthiness rate of 3.12 is reasonable since two users have rejected its friendship requests. The RDM method also considers the trustworthiness rates of the users who have accepted or rejected X's friendship requests.

3. The user represented by node X has sent three friendship requests to the users represented by nodes B, D, and E, who have all accepted the requests. In turn, X has also received two friendship requests from A and C.

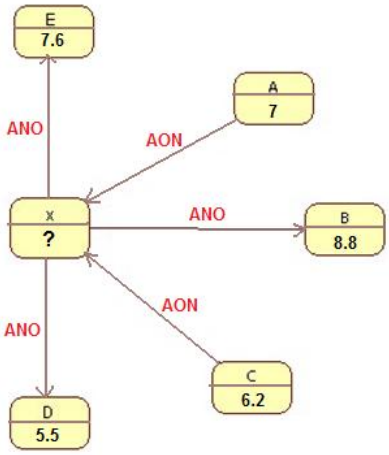

Figure 5. RDM method—fourth scenario

Based on the RDM method, X's trustworthiness rate = 8.3. Node X's trustworthiness rate of 8.3 is reasonable since X has received two friendship requests, and no friendship request sent by X

has been rejected. The method also considers the trustworthiness rates of the users who have sent friendship requests.

EXPERIMENTS AND VALIDATION

To validate the proposed method, two experiments have been conducted. Both of which have been conducted using an online simulation software (environment) that mimics Facebook. The simulation software that have been used in the first experiment was provided with a feature that evaluates and shows the trustworthiness of the users, based on the proposed method in this paper. However, the second experiment has been conducted using a simulation software without the feature that evaluates the trustworthiness of the users. The second experiment has been used as a control group in order to examine to what extent the proposed method can be useful. The subjects for both experiments were undergraduate students who have been chosen randomly from multiple universities located in Saudi Arabia, and they used Facebook for at least one year.

First Experiment

As explained earlier, the first experiment was provided with a feature that evaluates the trustworthiness of the users, based on the proposed method in this paper. Two groups of subjects have been used in the first experiment. The subjects in first group have been asked to create fake profiles (for simplicity we call them fake group) and the have been asked to try to influence other users to make friendship with them (sending them friendship invitation or accepting their friendship invitation). The subjects in the second group have been asked to create real profiles, with their real information, just as they appear in their real Facebook profiles (for simplicity we call them real group). The simulation software allows users to choose their information that can be seen by others such as names, number of friends, location, education, and so on.

In order to encourage subjects to participate and pay attention to the provided information, the subjects were offered an amount equivalent to US$1 for every 20 points that they earn. The subjects of both groups have been told that they will earn a point for every successful connection to a real user, and loose a point for every connection to a fake user. In addition, the subjects of the fake group have been asked to provide some specific information, which can be collected from real user's profiles, and they will earn two points for every asked information if they are provided successfully. The subjects of both groups have been told that the users who are available in the simulation software might be real or fake users. However, the instructions have been provided for every group without knowing about the detailed instructions for the other group. All subjects have been known to the research team before starting the experiment, every subject has been asked to create only one profile, and the creation of every profile can be validated only by the research team. The subjects have been told that the points will be calculated in a month period of time.

Second Experiment

The second experiment has been conducted using similar procedures to the first experiment. The only deference between the first and the second experiment is that the simulation software that has been used in the second experiment is provided without the feature that evaluates and shows the trustworthiness of the users. Conducting the second experiment helps showing the extent to which the proposed method, which evaluates the trustworthiness of the users, can help in identifying real profiles from fake profiles.

Experiments' Result

Table 1 shows the result of both conducted experiments after a month period of time. The result shows that the average number of friends (number of connections) that every subject in the fake group (potential attackers) is connected to is 47 for the software that does not include RTM feature, while it is only 7 for the software that includes RTM feature. This means that RTM helps users (potential victims) identify fake profiles (potential attackers) and therefore not sending them friendship invitations or not accepting friendship invitations from them.

The result shows also that average number of friends that every subject in the real group (legitimate users) is connected to is 26 for the software that does not include RTM feature, while it is 27 (22 of them are real) for the software that includes RTM feature. This perhaps shows that RTM helps in maintaining the healthy sociability and connections between real (legitimate) users, while it reduces the connections with fake profiles (potential attackers).

In addition, the result shows that the average number of friendship invitations that have been sent by subjects in the fake group is 2445 invitations for the software that does not include RTM feature, and 1493 of them have been accepted. On the other hand, the average number of friendship invitations that have been sent by subjects in the fake group for the software that includes RTM feature is only 604 invitations, and only 71 of them have been accepted. This means that RTM helps in reducing sending fake invitations, and more importantly reducing accepting such invitations. The fact that RTM takes rejected invitations into account when calculating trustworthiness of a user, made every user perhaps (including fake group) cautious when sending friendship invitations, as they know that rejecting such invitation will minimize their trustworthiness rate.

Interestingly, the result shows that RTM does not make a significant change in terms of the average number of friendship invitations that have been sent by subjects in the real group. However, the result shows that RTM has made a significant improvement in helping users recognize the real (legitimate) users, and therefore influence them to accept their friendship invitations. That is, for the software that does not include RTM feature, out of 364 invitations that have been sent by subjects in the real group, only 75 have been accepted. On the other hand, for the software that includes RTM feature, out of 302 invitations that have been sent by subjects in the real group, 276 have been accepted. This result supports the argument that RTM helps in maintaining the secured and safe sociability and connections between legitimate users.

Finally, the result shows that RTM helps in reducing the disclosure of personal identifiable information (PII). That is, 57 % of the information which has been asked from subject in the fake group has been provided successfully for the software that does not include RTM feature, while it is only 18 % for the software that include RTM feature. This reduction perhaps resulted from the reduction of the number of connections with real users.

Element	Software with trustworthiness evaluation (RTM) (First Experiment)	Software without trustworthiness evaluation (RTM) (Second Experiment)
Average number_of_friends that every subject in the fake group is connected to.	7 (2 real and 5 fake)	47 (25 real and 22 fake)
Average number_of_friends that every subject in the real group is connected to.	27 (22 real and 5 fake)	26 (11 real and 15 fake)
Average number of friendship invitations that have been sent by subjects in the fake group, and how many of them have been rejected.	604 (71 accepted, 533 rejected)	2445 (1493 accepted, 952 rejected)
Average number of friendship invitations that have been sent by subjects in the real group, and how many of them have been rejected.	302 (276 accepted, 26 rejected)	364 (75 accepted, 289 rejected)
Percentage of provided information, which has been asked from subject in the fake group.	18 % has been provided successfully	57 % has been provided successfully

Table 1. Experiments' result

DISCUSSION AND CONCLUSION

The SNS users have been found to be quite vulnerable to falling victim to many social engineering tricks and attacks, such as phishing, clickjacking, sexual abuse, financial abuse, identity theft, impersonation, physical crime, and many other forms. As explained, the simple trick of offering free mobile phone minutes accounted for the largest number of attacks on Facebook users in 2013, increasing from 56% in 2012 to 81% in 2013 (Mazzuca, 2014). The magnitude of the problem highlights this study's significant contribution to the SNS sector. Moreover, recent research on SNS security shows that most social engineering threats, such as spamming, identity cloning, and bots, rely on fake identities (Fire et al., 2014), and an estimated 83 million Facebook accounts may be fake (Couper, 2013). The lack of authentication, which requires only an email address to create a new account in most SNSs, allows attackers to create as many profiles as they want, including profiles with fake and impersonated identities. This situation reflects the high demand for research that authenticates users or estimates their trustworthiness. As such, this study makes a vital contribution by attempting to resolve a serious information security issue.

This study's findings enrich (directly or indirectly) the literature in several research areas, such as source credibility, deception, persuasion, and information security management (e.g., phishing). One of the most important and probably the most challenging issues is how to detect attackers based on their activities in SNSs. Solutions to this problem are valuable for social engineering in SNSs. However, it has been reported (e.g., Algarni, Xu, & Chan, 2015; Egele, Stringhini, Kruegel, & Vigna, 2013; Viswanath et al., 2014) that is very difficult to distinguish between the activities of legitimate users and those of attackers. The proposed method is perhaps one of the most effective methods of addressing such a challenge. In terms of user susceptibility, several studies on information systems have investigated individuals' predisposition to security victimization by studying employees' compliance with organizations' security policies. Such research has been conducted by relying on a number of theories and techniques, such as the protection motivation theory (e.g., Johnston, Warkentin, & Siponen, 2015; Posey, Roberts, Lowry, Bennett, & Courtney, 2013; Posey, Roberts, Lowry, & Hightower, 2014), electroencephalography (e.g., Vance, Anderson, Kirwan, & Eargle, 2014), the technology threat avoidance theory (Herath et al., 2014), and the routine activity theory (e.g., Wang, Gupta, & Raj, 2015). There seems to be a general agreement that an individual's vulnerability to an attacker's deceptions is associated with making inaccurate judgments regarding the source.

This study contributes to this stream of research by providing a mechanism that helps users make accurate judgments about the sources' credibility and trustworthiness.

The main strength of the proposed method is its simplicity. Unlike other proposed methods, which require heavy calculations and rely on many factors, such as similarity between users, profile content analysis, and complex behavioral analysis, RDM evaluates trustworthiness based on a simple yet reliable factor. Additionally, the method shows strength in taking into account the trustworthiness rates of the user's friends but not the number of connected friends. This point reduces the risk of an attacker creating fake profiles and making friends with himself or herself to earn a higher trustworthiness rate.

Another strength of the proposed method is that it considers the trustworthiness rates of the intended recipients who reject the sender's friendship requests, while it gives the sender just a fair (5) trustworthiness rate for accepting the his/her friendship requests. This point reduces the risk of an attacker sending many requests to other users in order to earn a higher trustworthiness rate. At the same time, this feature increases the likelihood of rejecting the sender's (e.g., attacker or spammer) friendship requests and therefore decreases the sender's trustworthiness rate.

It is noteworthy that the method shows a limitation regarding a new user, who has not yet sent or received any friendship request (a cold start situation). In other words, the equation of the proposed method evaluates the new user's trustworthiness rate as = 0, but the new user should (arguably) be given a fair rate (5). However, this drawback disappears after the first friendship request is sent or received. Therefore, this shortcoming can be easily eliminated by adjusting the algorithm to assign any new user a fair trustworthiness rate (5) until that user sends or receives a friendship request.

As every research has its shortcomings, future researchers can utilize other creative methodologies to avoid this study's limitation (as discussed in the preceding section). For example, it would be interesting to apply the proposed method in an actual data set that is obtained from SNSs, such as Facebook, to test its efficiency. Research on social engineering in SNSs is still limited, and many challenging issues should be addressed, such as identity theft, impersonations, cloning attacks, phishing, and clickjacking. The research in these areas is multidisciplinary; therefore, we highly recommend closer collaboration among the information ecology, data science, and information security disciplines.

References

Abbasi, M.-A., & Liu, H. (2013). Measuring user credibility in social media *Social Computing, Behavioral-Cultural Modeling and Prediction* (pp. 441-448): Springer.

Al Zamal, F., Liu, W., & Ruths, D. (2012). *Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors.* Paper presented at the ICWSM.

Algarni, A., Xu, Y., & Chan, T. (2015). Susceptibility to social engineering in social networking sites: The case of Facebook.

Algarni, A., Xu, Y., Chan, T., & Tian, Y.-C. (2013a). *Social engineering in social networking sites: Affect-based model.* Paper presented at the Internet Technology and Secured Transactions (ICITST), 2013 8th International Conference for.

Algarni, A., Xu, Y., Chan, T., & Tian, Y.-C. (2013b). Toward understanding social engineering. *Law & Practice: Critical Analysis and Legal Reasoning*, 279-300.

Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). *Detecting spammers on twitter.* Paper presented at the Collaboration, electronic messaging, anti-abuse and spam conference (CEAS).

Boorman, J., Liu, Y., Zhang, Y., Bai, Y., Yao, S., Wang, M., & Tai, L. (2014). Implications of social media networks on information security risks.

Braun, R., & Esswein, W. (2013). *Towards a Conceptualization of Corporate Risks in Online Social Networks: A Literature Based Overview of Risks.* Paper presented at the Enterprise Distributed Object Computing Conference (EDOC), 2013 17th IEEE International.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology, 3*(2), 77-101.

Brody, R. G. (2012). Flying under the radar: social engineering. *International Journal of Accounting and Information Management, 20*(4), 335-347. doi: 10.1108/18347641211272731

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). *Discriminating gender on Twitter.* Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Castillo, C., Mendoza, M., & Poblete, B. (2011). *Information credibility on twitter.* Paper presented at the Proceedings of the 20th international conference on World wide web.

Chitrey, A., Singh, D., & Singh, V. (2012). A Comprehensive Study of Social Engineering Based Attacks in India to Develop a Conceptual Model. *International Journal of Information and Network Security (IJINS), 1*(2), 45-53.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *Dependable and Secure Computing, IEEE Transactions on, 9*(6), 811-824.

Conti, M., Poovendran, R., & Secchiero, M. (2012). *FakeBook: detecting fake profiles in on-line social networks.* Paper presented at the Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012).

Couper, M. (2013). *Is the sky falling? New technology, changing media, and the future of surveys.* Paper presented at the Survey Research Methods.

Denzin, N. K., & Lincoln, Y. S. (2005). *The Sage handbook of qualitative research*: Sage.

Dimensional-Research. (2011). The risk of social engineering on information security: a survey of it professionals. Technical Report, Long Beach, CA.

Egele, M., Stringhini, G., Kruegel, C., & Vigna, G. (2013). *Compa: Detecting compromised accounts on social networks.* Paper presented at the NDSS.

Eisend, M. (2006). Source credibility dimensions in marketing communication–A generalized solution. *Journal of Empirical Generalizations in Marketing, 10*(2), 1-33.

Fire, M., Goldschmidt, R., & Elovici, Y. (2014). Online Social Networks: Threats and Solutions. *Communications Surveys & Tutorials, IEEE, 16*(4), 2019-2036.

Fire, M., Katz, G., & Elovici, Y. (2012). Strangers intrusion detection-detecting spammers and fake proles in social networks based on topology anomalies. *HUMAN, 1*(1), pp. 26-39.

Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., & Zhao, B. Y. (2010). *Detecting and characterizing social spam campaigns.* Paper presented at the Proceedings of the 10th ACM SIGCOMM conference on Internet measurement.

Hadnagy, C. (2010). *Social engineering: The art of human hacking*: Wiley.

Herath, T., Chen, R., Wang, J., Banjara, K., Wilbur, J., & Rao, H. R. (2014). Security services as coping mechanisms: an investigation into user intention to adopt an email authentication service. *Information systems journal, 24*(1), 61-84.

Hogben, G. (2007). Security issues and recommendations for online social networks. *ENISA position paper, 1*.

Huber, M., Kowalski, S., Nohlberg, M., & Tjoa, S. (2009). *Towards automating social engineering using social networking sites.* Paper presented at the Computational Science and Engineering, 2009. CSE'09. International Conference on.

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM, 50*(10), 94-100. doi: 10.1145/1290958.1290968

Johnston, A. C., Warkentin, M., & Siponen, M. (2015). An enhanced fear appeal rhetorical framework: Leveraging threats to the human asset through sanctioning rhetoric. *Mis Quarterly, 39*(1), 113-134.

Krombholz, K., Hobel, H., Huber, M., & Weippl, E. (2014). Advanced social engineering attacks. *Journal of Information Security and Applications*.

Laura Mazzuca, T. (2014). 7 Scary Findings from the 2014 Symantec Internet Security Threat Report. *Property & Casualty 360*.

Liu, W., & Ruths, D. (2013). *What's in a Name? Using First Names as Features for Gender Inference in Twitter.* Paper presented at the AAAI Spring Symposium: Analyzing Microtext.

Mazzuca, T. (2014). 7 Scary Findings from the 2014 Symantec Internet Security Threat Report. *Property & Casualty 360*.

McCord, M., & Chuah, M. (2011). Spam detection on twitter using traditional classifiers *Autonomic and Trusted Computing* (pp. 175-186): Springer.

Meligy, A. M., Ibrahim, H. M., & Torky, M. F. (2015). A Framework for Detecting Cloning Attacks in OSN Based on a Novel Social Graph Topology.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *ICWSM, 11*, 5th.

Nagy, J., & Pecho, P. (2009). Social Networks Security. 321-325. doi: 10.1109/securware.2009.56

Nohlberg, M. (2009). Why Humans are the weakest Link. *Social and Human Elements of Information Security: Emerging Trends*.

Posey, C., Roberts, T., Lowry, P. B., Bennett, B., & Courtney, J. (2013). Insiders' protection of organizational information assets: Development of a systematics-based taxonomy and theory of diversity for protection-motivated behaviors. *Mis Quarterly, 37*(4), 1189-1210.

Posey, C., Roberts, T. L., Lowry, P. B., & Hightower, R. T. (2014). Bridging the divide: A qualitative comparison of information security thought patterns between information security professionals and ordinary organizational insiders. *Information & Management, 51*(5), 551-567.

Rahman, M. S., Huang, T.-K., Madhyastha, H. V., & Faloutsos, M. (2012). *Efficient and Scalable Socware Detection in Online Social Networks.* Paper presented at the USENIX Security Symposium.

Rao, D., Paul, M. J., Fink, C., Yarowsky, D., Oates, T., & Coppersmith, G. (2011). Hierarchical Bayesian Models for Latent Attribute Detection in Social Media. *ICWSM, 11*, 598-601.

Roberts, K. L., & Taylor, B. (2002). *Nursing research processes: An Australian perspective*: Nelson.

Shariff, S. M., & Zhang, X. (2014). *A survey on deceptions in online social networks.* Paper presented at the Computer and Information Sciences (ICCOINS), 2014 International Conference on.

Stringhini, G., Kruegel, C., & Vigna, G. (2010a). *Detecting spammers on social networks.* Paper presented at the Proceedings of the 26th Annual Computer Security Applications Conference.

Stringhini, G., Kruegel, C., & Vigna, G. (2010b). A study on social network spam. *GSWC 2010*, 43.

Thomas, K., McCoy, D., Grier, C., Kolcz, A., & Paxson, V. (2013). *Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse.* Paper presented at the USENIX Security.

Thornburgh, T. (2004). *Social engineering: the dark art.* Paper presented at the Proceedings of the 1st annual conference on Information security curriculum development.

Vance, A., Anderson, B. B., Kirwan, C. B., & Eargle, D. (2014). Using measures of risk perception to predict information security behavior: Insights from electroencephalography (EEG). *J. Assoc. Inf. Syst, 15*(10), 679-722.

Viswanath, B., Bashir, M. A., Crovella, M., Guha, S., Gummadi, K. P., Krishnamurthy, B., & Mislove, A. (2014). *Towards Detecting Anomalous User Behavior in Online Social Networks.* Paper presented at the USENIX Security Symposium.

Wagner, C., Mitter, S., Körner, C., & Strohmaier, M. (2012). When social bots attack: Modeling susceptibility of users in online social networks. *Making Sense of Microposts (# MSM2012)*, 2.

Wang, A. H. (2010). *Don't follow me: Spam detection in twitter.* Paper presented at the Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on.

Wang, J., Gupta, M., & Raj, R. (2015). Insider Threats in a Financial Institution: Analysis of Attack-Proneness of Information Systems Applications. *Management Information Systems Quarterly, 39*(1), 91-112.

West, R., Mayhorn, C., Hardee, J., & Mendel, J. (2009). The Weakest Link: A Psychological Perspective on Why. *Social and Human Elements of Information Security: Emerging Trends*.

Workman, M. (2007). Gaining access with social engineering: An empirical study of the threat. *Information Systems Security, 16*(6), 315-331.

Yardi, S., Romero, D., & Schoenebeck, G. (2009). Detecting spam in a twitter network. *First Monday, 15*(1).

