Understanding How Big Data Leads to Social Networking Vulnerability

Romany F. Mansour

Abstract Although the term "Big Data" is often used to refer to large datasets generated by science and engineering or business analytics efforts, increasingly it is used to refer to social networking websites and the enormous quantities of personal information, posts, and networking activities contained therein. The quantity and sensitive nature of this information constitutes both a fascinating means of inferring sociological parameters and a grave risk for security of privacy. The present study aimed to find evidence in the literature that malware has already adapted, to a significant degree, to this specific form of Big Data. Evidence of the potential for abuse of personal information was found: predictive models for personal traits of Facebook users are alarmingly effective with only a minimal depth of information, "Likes". It is likely that more complex forms of information (e.g. posts, photos, connections, statuses) could lead to an unprecedented level of intrusiveness and familiarity with sensitive personal information. Support for the view that this potential for abuse of private information is being exploited was found in research describing the rapid adaptation of malware to social networking sites, for the purposes of social engineering and involuntary surrendering of personal information.

1 Introduction

Exactly how much can be known from a user's online social networking profile or profiles? These days, more and more people are spending significant portions of time every day on social networking. In 2011, the worldwide average for Facebook was 40 min for 800 million users, according to Los Angeles Times (2011). In fact, the sheer quantity of social interaction now occurring over social networking is such that a qualitative shift is taking place in our globalized society. This shift is

R. F. Mansour (✉)
Faculty of Science, Department of Mathematics, New Valley,
Assiut University, Asyut, Egypt
e-mail: romanyf@aun.edu.eg

S. S. Roy et al. (eds.), *Big Data in Engineering Applications*,
Studies in Big Data 44, https://doi.org/10.1007/978-981-10-8476-8_10

towards a replacement, in many ways, of in-person social interaction with interaction over social networking [1]. For example, social rituals or rites such as deciding whether a person would be suitable for dating now often occur first over Facebook or other social networking sites. There is a rise in the studies in social networking through its development, its effect to the global economy and the human psychology behind the use of these social networks [2]. Employers are also likely to screen prospective employees through an examination of their social networking profiles, especially Facebook and LinkedIn. The iniquitousness of social networking websites makes them immense repositories of personal information, carrying grave risks for abuse of privacy at the hands of malware. It is important that malware that depends on such Big Data techniques to perform social engineering and other unethical or socially compromising activities be more fully identified, characterized, and ultimately addressed.

Objectives of the study

I. To find out how much personal information can be obtained from the social networking sites.
II. To find out the privacy risks associated with personal information on social networking sites.

Significance of the study

Understanding social networking is an important aspect for users. This study will help the users identify the risks that are associated with the exposure of their personal information and how well they can mitigate these risks. Maintaining privacy of the users in the social networks is a necessary agenda for the users of the social networks.

2 Methods

The literature was examined for two separate lines of evidence related to the risk of dire loss of privacy as a result of Big Data—based mining of social networking website information. First, literature dealing with the theoretical potential for inferring personal details of users of social networking websites was searched for. Searches were performed on Google Scholar and Web of Science, using the terms "social networking", "social engineering", "big data", and "predictive models".

The second line of literature research aimed to discover evidence the malware is already adapting to exploit the potential of social networking websites and degrading privacy of users. Again, Google Scholar and Web of Science were used. However, in this case, the search terms were extended to include "malware", "phishing", and "hacking".

For both lines in literature research and inquiry, only articles from the last 5 years (2010–2015) were considered.

3 Results

3.1 Potential of Big Data Techniques for the Inference of Sensitive Personal Information

The study by [3] used six different features of a sizeable sample of 180,000 Facebook users' profiles to predict personality traits. The personality trait measurement method used was the standard Five Factor Model, which measures the level of the following personality traits: Extraversion, Neuroticism, Agreeableness, Openness, and Conscientiousness. The six features used by [3], summarized in Table 1, are numbers of: Facebook friends, associations with groups, Facebook "likes", photos uploaded by user, status updates by users, and times others "tagged" user in photos. The 180,000 volunteers who provided information from their facebook profiles also completed the Five Factor Model personality test. Therefore, it was possible to compare predictions from the Facebook model to objective results from the Five Factor Model. Using multiple regression, the authors found that predictions from the Facebook model could be generated that were very accurate, assuming that the results from the Five Factor Model did not incorporate any misrepresentations of personality. These findings supported findings from an earlier work that social networking profiles do not present an idealized or skewed version of a user's persona, but rather a realistic and fairly objective summary [4]. The [3] study did find, however, that the traits of "Agreeableness" and "Openness" were significantly ($p < 0.05$) less accurately predicted than were the other three traits. A somewhat later, but similar, study reported the ability to predict personality traits using a natural-language parsing model to automatically analyze individuals' statuses [5]. This model was trained on a corpus of over 700 essays that had been manually curated and assigned labels with the appropriate amounts of the five favors (Openness, Agreeableness, Extraversion, Neuroticism, and Conscientiousness) assigned. This study corroborated the findings of the [3] study that personality traits could be accurately inferred.

Perhaps the most recent transformative research on the subject of inferring personal details from facebook or other social networking information was reported by [6]. This group took a sample of 58,000 volunteers who had made part of their Facebook information available (Facebook 'Likes"). The authors were able to show that a list of a person's likes, which are highly visible as they are generally

Table 1 Features used by [3] to predict Facebook user personality traits (according to Five Factor Model)

Feature	Details
Friends	Number of Facebook friends
Groups	Number of associations with groups
Likes	Number of Facebook "likes"
Photos	Number of photos uploaded by user
Statuses	Number of status updates by user
Tags	Number of times others "tagged" user in photos

publically available, can be used to predict certain demographic and personal pieces of information with great accuracy. The categories of personal information that were predicted were diverse, but among those that could be predicted with high accuracy were sexual orientation, ethnicity, religion, political orientation, personality, IQ, drug use and various other pieces of personal and family information. The most accurately predicted demographic and personal factors were sexual orientation in men (88%), African American versus Caucasian American (95%), and political orientation (Democrat or Republican) (85%). Thus, a large amount of personal information of great relevance to potential employers can be predicted from an individual's collection of "Likes" on Facebook [4, 7], argue that such information on the web can be used in carrying bout advertisements targeting a specific group of people. Social networks can provide useful information about the users that can be useful to the marketers in laying down their marketing strategies.

Jernigan and Mistree [8], carried out a study among MIT students based on the hypothesis that the number of an individual's Facebook friends can be used to determine the sexual orientation group of the user. A thorough analysis was carried out on the students who used the MIT browsers. The study revealed that the number of friends that an individual has can be used to predict the sex orientation of the user. For instance is a user has more homosexual friends then the likeliness that the individual is homosexual is very high. The findings are summarized in the Table 2.

It has also been found that people's social strategies, and therefore possibly even the underlying social motivations, can be inferred from a careful analysis of Facebook and social networking patterns. For example, through an analysis of the evolution of Facebook connections over time, [9] were able to differentiate non-social capital seeking from social capital seeking friends. The researchers developed a predictive model based on the patterns of connectivity over time, and found that these patterns only differed significantly from normal when an individual was making connections with the intentional goal of seeking social capital. For

Table 2 Percentage friends per sex orientation group

Sex orientation group	Percentage friends per group					
	Heterosexual (%)		Bisexual (%)		Homosexual (%)	
Heterosexual						
Female	19.0	22.4	0.7	0.5	0.4	0.8
Male	13.9	28.3	0.5	0.4	0.3	0.7
Bisexual						
Female	15.5	20.7	1.4	1.1	0.3	1.2
Male	12.6	22.3	0.8	0.6	0.3	1.9
Homosexual						
Female	18.0	23.6	0.9	0.7	0.2	0.8
Male	13.1	21.4	1.1	1.1	0.4	4.6

Retrieved from: [8]

example, if an individual has recently been introduced to a new group, he or she is likely to first connect with a central hub in the Facebook environment for the group of people, and then rapidly add connections (which then become mutual connections between the individual and the central hub). This central hub is often someone in a position of power or privilege. Not only can analysis of a person's friend connection patterns reveal social intent, but it can reveal who a person's real friends are more likely to be, and who a person has "friended" merely as acquaintances. In a reversal of the predictive methodology, [10] used measured traits of personality to predict Facebook usage. Specifically, the researchers were able to find that certain personality traits (neuroticism foremost) were strong predictors of wall posting "regret", or the tendency to remove a posting on a user's own, or a friend's wall. Hughes et al. reviewed work done to create predictors with data from Facebook versus from Twitter, finding the two sites to be very similar overall.

Ross et al. [1], carried out a similar study on how personality traits and competency influenced the way in which university students utilized Facebook for social purposes and came up with different results centrally to the ones discussed above. The research utilized 97 students from the Southwestern Ontario as the respondents, in which 85 were women and 25 were men. A 28 item questionnaire was used as the study tool. The authors found out that some personality characteristics influenced the Facebook use but their level on the impact of Facebook use differed greatly. The students who scored highly in the extraversion characteristics belonged to more Facebook groups but had very few friends. The reason behind this is that some of the users prefer instant contact with the friends a feature which is not enabled with the Facebook. Therefore they choose not to use Facebook as their primary source of interaction. Those high in the neuroticism character preferred their walls as compared to those low in the same personality as they preferred photos. Those who preferred wall posting are associated with the ability to think out well before positing. Those who post photos are exposed to privacy intrusion as such photos contain some personal information such as place where the photo was posted from. Openness and experience in utilization of Facebook was associated with the ability to understand on how to use the several elements if Facebook, how to comment and how to use other Facebook feature. However more agreeable individuals contained lesser online contacts, also there was no significant relationship between conscientiousness with the utilization of Facebook.

In their study [11] carried out a comparison between the social culture of interaction between the two social virtual world of China, Uworld and HiPiHi. Unlike Uworld, HiPiHi makes the use of the social networking to promote its business products. This has become another strategy for promoting business activities through the virtual games online. On the other hand Uworld provides entertainment games which are in different forums that are not related to businesses. In these forums. In the Uworld the users can make and chat with friends in the virtual room and also playing games. This makes it possible to create more friends in the Uworld than in the HiPiHi.

In summary, it appears that there is currently a surprising amount of information that can be inferred from a user's social networking profile. As emphasized by [12],

people social agendas could be revealed. Indeed, in some cases, it seems possible that models based on Facebook "Likes", for instance, might be able to correctly make predictions that an individual himself would never have known. This is possible thanks to the vast sample size available (nearly a billion users worldwide, just for Facebook), as well as the richness, standardization, and quantity of information that is routinely deposited on Facebook by users. One cannot help but speculate that this diversity and potency of information could be used to intelligently craft tools and traps to manipulate users of social networking websites, or indeed, other websites (after saving information from the users' profiles).

3.2 Social Networking Sites and Malware Risk

Social engineering occurs perhaps most directly on websites where malware and phishing programs are able to induce internet browsers and users into places where security is less available or effective. Quite often, bright-colored ads or links artificially placed high in the results from search engines lead users stray into areas where their ability to detect malware is reduced [13], as a result of a weaker firewall, less visible pop-ups, or the leverage of anti-anti-malware tools.

Tracking the behavior and attack styles of these socially-engineering forms of malware could be a very interesting and compelling, modern and promising way to go about thesis research. A recent article collected information on the intensity and frequency of malware [14]. This article found, for example, that the pervasiveness of malware is generally due to the use of common avenues of attack. The group further found that such malware relies on two primary strategies, technological and psychological manipulation. Technological manipulation includes placing fake versions of functional navigational buttons over the actual buttons on the graphical user interface of social networking websites, or having the link pop up the instant the user clicks. Psychological manipulations involve listing unsponsored pop-ups in the side bar that supposed the user finds appealing enough to want to click on, regardless of prior plans on the website. The advantage, from the perspective of the malware, of hijacking personal information on social networking websites, is that users are often rather less rushed or focused in their browsing habits, and therefore can be more easily led astray [14]. Further elucidating sub-types of these two primary types of social engineering (technological and psychological manipulation) could be a compelling goal for thesis research. The limiting factor in this case might be access to sufficient user profiles, and the resistance one would likely encounter when trying to avail oneself of the user profiles when the users are informed that a virus is to be run on their system or targeted at their user profile.

A number of other areas exist in the internet wherein fraud in its various forms takes place. In general, whenever a great deal of technological competence is required, it becomes easier for malware to defraud an individual by false or alternative navigation around the website(s). In general, any area or circumstance in which the individual is suddenly faced with a request or demand seeming to

emanate from a technically-knowledge authority are far more likely than average to lead to incidences of internet fraud [15].

3.3 Social Engineering on Social Networking Sites

Social networking sites are particularly prone to unknowingly or unwillingly giving a platform for the attack of such malware. Social networking sites are some of the biggest and most popular, and although incredible amounts of data exist, the study of social networking is still in its infancy. Because most social-engineering types of viruses are found on social networking sites, it would be fairly direct and intuitive to design a thesis around the habits of users who fall victim to more malware (or to generate and provide evidence for/against other hypotheses [13]. This malware could take a number of forms, as the information on Facebook is sufficient, for nearly all individuals, to infer a great deal of additional very personal and sensitive information.

Not only are social websites ideal for leading user astray, but by virtue of their sheer size and versatile functionality these websites also contain unprecedented amounts of valuable personal information. Even if such personal information is not directly provided by the user on the website himself or herself, it may still be obtainable for malware, by dint of tunneling through privacy restrictions and reading, e.g., information from instant message conversations [16]. Through these conversations unauthentic messages can be sent. Often, these messages are not obviously "robotic" in nature, but rather have greetings from supposed people (users on facebook) as their first line of attack to disarm and socially position the victim for further information attacks. "Bots", for example, may replicate themselves and even generate false pictures and histories, and by first friending a victim and then posting indirectly related material, induce the victim to actually make first contact and assume himself/herself to be in charge of the social situation. In fact, this trust and "belief" in the legitimacy of the communication disarms the user, compelling him or her to surrender valuable personal information or even money.

4 Conclusion and Discussion

Even without soliciting information directly from a user of a social networking site, hackers, malware distributors, or other internet social engineers could quite easily infer a great deal of personal information from users, based simply on the users' profiles and networking behavior. The potential for abuse is clear—[3] show that analysis of profile information about Facebook users at the Big Data level (thousands of users) can lead to profiling of personal characteristics across a broad range

of factors. More profoundly, [6] find that just using Facebook "likes" allows for the creation of predictive models that indicate an individual's range, sexual orientation, and other sensitive demographic and personal details with alarming accuracy, up to 98% in the case of race. Undoubtedly, models can only be made stronger with the addition of more complex and rich data, e.g. from the mining of status updates, history, social connections, groups, and even pictures. Facebook is already capable of identifying facial features and other features of environs presented in photos.

Equally importantly, it is clear that bots and malware have already evolved that take advantage of the social milieu and at least some personal details of users to lead users astray, e.g. into less secure sites where further personal information can be stripped away. These bots and malware take advantage of the high level of activity the users engage in, when navigating through social networking sites. Mimicry of more legitimate ads targeted to users makes malware difficult to spot, especially for a distracted and enthusiastic user. It is important that these trends are recognized and reversed, before they can become even more powerful and insidious.

4.1 Recommendations

Enhancing privacy settings is a key strategies in mitigating privacy risks in the social networks. Setting privacy settings and cookies that can detect malwares and block them automatically help in dealing with vulnerabilities in social networks [17]. Authentication mechanisms can also be used to avoid hijackers or non-authorized users from login in into an individual's account [18]. The operators have also provided internal protection mechanisms that protect and detect spams or other such messages which are designed to collect user's personal information secretly [19]. Commercial solutions too can work by purchasing specialized softwares that have ability to defend user against any form of cyber-attacks [20].

4.2 Future Research

The need for personal security in the social networks has become increasingly important. Several solutions have been suggested and implemented but still the problem persists and several people have lost a lot of their resources due to these attacks. There is need to carry out research on how effective the adopted solutions are in helping solve these problems in the ever dynamic field of technology and need for coming up with new strategies of solving the problem.

References

1. Ross, C., Orr, E. S., Sisic, M., Arseneault, J. M., Simmering, M. G., & Orr, R. R. (2009). Personality and motivations associated with Facebook use. *Computers in Human Behavior, 25*(2), 578–586.
2. Zhang, X., Wang, W., de Pablos, P., Tang, J., & Yan, X. (2015). Mapping development of social media research through different disciplines: Collaborative learning in management and computer science. *Computers in Human Behaviour, 51,* 1142–1153.
3. Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., & Stillwell, D. (2012, June). Personality and patterns of Facebook usage. In *Proceedings of the 3rd Annual ACM Web Science Conference* (pp. 24–32). ACM.
4. Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., et al. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science, 21* (3), 372–374.
5. Farnadi, G., Zoghbi, S., Moens, M. F., & De Cock, M. (2013). How well do your Facebook status updates express your personality? In *Proceedings of the 22nd Edition of the Annual Belgian-Dutch Conference on Machine Learning (BENELEARN)*.
6. Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*(15), 5802–5805.
7. De Bock, K., & Van Den Poel, D. (2010). Predicting website audience demographics for Web advertising targeting using multi-website clickstream data. *Fundamenta Informaticae, 98*(1), 49–70.
8. Jernigan, C., & Mistree, B. F. (2009). Gaydar: Facebook friendships expose sexual orientation. *First Monday*, *14*(10).
9. Ellison, N. B., Steinfield, C., & Lampe, C. (2011). Connection strategies: Social capital implications of Facebook-enabled communication practices. *New Media & Society, 13*(6), 873–892.
10. Moore, K., & McElroy, J. C. (2012). The influence of personality on Facebook usage, wall postings, and regret. *Computers in Human Behavior, 28,* 267–274.
11. Zhang, X., de Pablos, P., Wang, X., Wang, W., & Sun, Y. (2014). Understanding the users' continuous adoption of 3D social virtual World in China: A comparative case study. *Computers in Human Behaviour, 35,* 578–585.
12. Butler, D. (2007). Data sharing threatens privacy. *Nature, 449*(7163), 644–645.
13. Algarni, A., Xu, Y., Chan, T., & Tian, Y.-C. (2013). Social engineering in social networking sites: Affect-based model. In *Proceedings of the 8th IEEE International Conference for Internet Technology and Secured Transactions (ICITST-2013)* (pp. 508–515). London: The Institute of Electrical and Electronics Engineering, Inc.
14. Abraham, S., & Chengalur-Smith. (2010, August). An overview of social engineering malware: Trends, tactics, and implications. *Technology in Society, 32*(3), 183–196.
15. Rusch, J. J. (1999). *The "social engineering" of Internet fraud*. USA: United States Department of Justice.
16. Laszka, A., Felegyhazi, M., & Buttyan, L. (2014). A survey of interdependent information security games. *ACM Computing Surveys (CSUR), 47*(2), 23.
17. Tipton, H. F., & Krause, M. (2012). *Information security management handbook*. CRC Press.
18. Whitman, M., & Mattord, H. (2011). *Principles of information security*. Cengage Learning.
19. Rasool, M. A., & Jamal, A. (2011). *Quality of freeware antivirus software.*
20. Sukwong, O., Kim, H. S., & Hoe, J. C. (2011). Commercial antivirus software effectiveness: An empirical study. *Computer, 44*(3), 0063–70.