Computers in Human Behavior 57 (2016) 348–351

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Understanding how big data leads to social networking vulnerability

Romany F. Mansour

Faculty of Science, N.V. Assiut University, Egypt

ARTICLE INFO

Article history:
Received 28 August 2015
Received in revised form
28 November 2015
Accepted 20 December 2015
Available online 31 December 2015

Keywords:
Big data
Social networking
Social engineering
Predictive models

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 towards a replacement, in many ways, of in-person social interaction with interaction over social networking (Ross et al., 2009). 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 (Zhang, Wang, de Pablos, Tang, & Yan, 2015). Employers are also likely to screen prospective employees through an examination of their social networking profiles, especially Facebook and LinkedIn. The ubiquitousness 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.

1.1. 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.

1.2. 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

E-mail address: romanyf@aun.edu.eg.

http://dx.doi.org/10.1016/j.chb.2015.12.055
0747-5632/

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 Bachrach, Kosinski, Graepel, Kohli, and Stillwell (2012) 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 Bachrach et al., 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 (Back et al., 2010). The Bachrach et al. (2012) study did find, however, that the traits of “Agreeableness” and “Openness” were significantly ($p < .05$) 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 (Farnadi, Zoghbi, Moens, & De Cock, 2013). 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 Bachrach et al. (2012) study

Table 1
Features used by Bachrach et al. (2012) 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 other “tagged” user in photos

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: Jernigan and Mistree (2009).

that personality traits could be accurately inferred (see Table 2).

Perhaps the most recent transformative research on the subject of inferring personal details from Facebook or other social networking information was reported by Kosinski, Stillwell, and Graepel (2013). 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 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 vs. 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 (Back et al. 2010). De Bock and Van Den Poel (2010), 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 (2009), 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 below.

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, (Ellison, Steinfield, & Lampe, 2011) 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 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