World Journal of Computer Application and Technology 1(2): 29-34, 2013 http://www.hrpub.org
DOI: 10.13189/wjcat.2013.010202

A Malicious Use of a Clustering Algorithm to Threaten the Privacy of a Social Networking Site User

Yeslam Al-Saggaf[1,*], Md Zahidul Islam[2]

[1]School of Computing and Mathematics, Charles Sturt University, Locked Bag 588, Boorooma Street, Wagga Wagga, NSW 2678, Australia
[2]Centre for Research in Complex Systems, School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW 2795, Australia
*Corresponding Author: yalsaggaf@csu.edu.au

Abstract This article explores the potential of data mining as a technique that could be used by malicious data miners to threaten the privacy of Social Network Sites (SNSs) users. It applies a data mining algorithm, specifically a clustering algorithm, to a hypothetical dataset to show the ease at which characteristics about the SNSs users can be discovered and used in a way that could invade their privacy. One important outcome of exploring the threats from data mining on individuals' privacy is enable SNSs developers better understand the ways in which SNSs can be used by malevolent data miners to harm users. Another important outcome is to help developers of SNSs develop mechanisms that will provide protection to users from these knowledge discovery techniques.

Keywords Data Mining, Social Network Sites, Privacy

1. Introduction

The web environment has changed dramatically in the last few years with the emergence of one of the most fertile environments of personal data on the web; social network sites (SNSs). A social network site (SNS), defined formally, is a web-based service that allows individuals to "1) construct a public or semi-public profile within a bounded system, 2) articulate a list of other users with whom they share a connection, and 3) view and traverse their list of connections and those made by others within the system"[1, p. 211]. Users are also prompted to invite other SNSs users to view and follow their personal profile and web behaviour within the SNSs and to label and categorise these relationships with other users as "friends" (Facebook) or sometimes "followers" (Twitter).

With hundreds of millions of web users around the world now uploading masses of personal data onto the web within the confines of personal networks of friends, family, and acquaintances, the success of social network sites, such as Facebook and Twitter, has made a wealth of social networking data available for businesses and more malicious data miners to profit from. Major culprits include banks and credit agencies, and advertising companies, who are usually assisted in their data mining practices by SNSs with whom they are affiliated.

Sar & Al-Saggaf [2] study has found that by just visiting non SNSs and SNSs, third party sites can track the users' activities by the use of the HTTP cookies. These third party sites can be advertising sites or data aggregators, or other SNSs. However, these sites are not only collecting these massive amounts of data about SNSs users; they are also mining them, for the purpose of placing the users in new and non obvious classifications or categories that the users themselves did not know they ever existed [3]. Tavani [3] further argues that the problem is that the current privacy laws do not offer individuals any protection with regards to how information about these users obtained through data mining are subsequently used, particularly to make decisions about them in light of the categories these data mining activities discover.

Indeed, individuals in Australia, for example, are not implicated by the privacy law [1]. In other words, the privacy laws don't implicate the public in their individual capacity. For example, the laws do not protect an individual's privacy if another individual breaches it. While individuals can be taken to court and tried in relation to privacy offenses, but only under, for example, breaches of confidence, or defamation laws, i.e. not under privacy laws. Privacy laws subject only businesses that have a turnover of three million dollars and over, which means even 95% of Australian businesses are not bound by these laws[2].

This raises an important question: what stops third party

[1] Australian Government office of the Australian Information professional. 2012. Business and Me. Retrieved march 9, 2012 from http://www.privacy.gov.au/individuals/business
[2] Personal communication with The Victorian privacy commissioner, Ms Helen Versey, on the 13 of February 2012 during the sixth Australian Institute of Computer ethics conference in Melbourne. The Victorian privacy commissioner gave the key note address at this conference.

sites from mining the data they acquire from SNSs and non SNSs and using the results of data mining activities to identify users or link them with groups they don't necessarily belong to and in ways that can threaten their privacy? There are many accounts of privacy in the literature but for the purpose of this article the Informational Privacy theory definition will be used. According to this theory, privacy is defined as the ability to restrict others access to and control over the flow of one's personal information, including the transfer and exchange of that information [3, pp. 136-137].

The aim of this article is to explore the potential of data mining as a technique that could be used by malicious data miners to threaten the privacy of SNSs users. The aim will be achieved by applying a data mining algorithm, specifically a clustering algorithm, to a hypothetical dataset to show the ease at which characteristics about the SNSs users can be discovered and used in a way that could invade their privacy. One important outcome of exploring the threats from data mining on individuals' privacy is enabling SNSs developers to better understand the ways in which SNSs can be used by malevolent data miners to harm users. Another important outcome is helping developers of SNSs develop mechanisms that will provide protection to users from these knowledge discovery techniques.

Before demonstrating how data mining can actually threaten the privacy of SNSs users, which is something that to the best of our knowledge of the literature has not been done before, the article will first provide a short introduction about data mining and will then review the literature in relation to some examples of methodologies employed in social network data collection and analysis. The reason for the inclusion of this section is to show that while this article used a hypothetical dataset, it is easy to build a dataset from an SNS.

2. Data Mining in SNSs

Data mining has traditionally been restricted to trolling through offline 'data warehouses', large storehouses of personal data consisting mainly of commercial transactions [4]. However, a profitable application of data mining, especially in recent years, has been in all sorts of web-based services. The World Wide Web provides web data miners with a virtually limitless repository of information on which to employ their data gathering techniques. Vast amounts of personal information (e.g. names, addresses, web browsing behaviour, personal preferences, online consumer records, etc) are made publically searchable through the content of web pages. Web data mining techniques can be employed to automatically discover, extract, and analyse information from web documents and services to generate new and financially lucrative data [5, 6].

There are three types of web data mining. The first type is web structure mining, which allows miners to extract the links and structure of websites for the purpose of discovering previously unknown relationships between the mined websites. Applying this technique results in a structural summary of the websites or web pages [5]. The second type is text mining, known also as web content mining [6]. This technique involves the scanning and extracting of contents from webpage, i.e. text, pictures, graphs, audio, video, and hyperlinks, for the purpose of determining the relevance of the contents to a search query [6]. One of the main uses of this method is gathering and summarising the best possible information available from web content for a user requesting that information [5]. The third type is web usage mining [5], which involves tracking and analysing the navigation behaviour of web users for the purpose of understanding the movements of users online and what they are doing [7]. One of the main uses of this method is to generate data that represents, without identifying the individual, the web users' actions in a public environment [7].

Web usage mining can identify traces such as an IP-address of a computer, type of browser used, date and time of login to a site, hyperlinks followed, use of cookies with frequent visits to particular sites, etc, can be extracted using web usage mining techniques to discover patterns in web browsing behaviour. Web usage mining has the potential to be privacy violating due to the fact that all internet service providers (ISPs) can link a web users' navigation behaviour to their personal information, which the user has had to provide to their ISP in order for the ISP to supply the user access to the internet [7].

3. Methodologies Used and Results Reported in SNSs Literature

The following summarises some examples of methodologies employed in social network data collection and analysis reported in the literature, as well as the results of these studies. While this article used a hypothetical dataset this section will show that it is easy to build a dataset from an SNS. Catanese et al [8] conducted an analysis of Facebook using a crawling data collection technique. Two different methodologies were employed to gather partial data, or a "sub-graph", of the complete Facebook structure. The reason for this was to test the hypothesis that it is possible to study an SNS without having access to its complete dataset. Two different sampling techniques were employed independently in the data collection process: a 'Breadth-first-search' crawler, and a 'Uniform' crawler. Despite the different techniques implemented, the same data collection processes was followed in each case. This consisted of: 1) preparing for the implementation of the crawler; 2) starting the process of data extraction; 3) the crawler then extracts lists of friends from Facebook on a repeating cycle; 4) raw data is collected until the extraction process finishes its processing (i.e. the algorithm visits all discovered nodes) or is manually stopped; 5) the raw data is then cleaned and duplication of information is discarded; 6) a social network graph is generated based on the data structure.

Relevant graphs were generated, with users represented by 'nodes' of the graphs, and the relationships among users represented by the edges of the graphs. The following properties were analysed based on the graphs generated: 1) the 'degree distribution' reflected in the graphs indicated a common structure in networks where a comparatively small amount of nodes (users) are connected through a large number of relationships; 2) the 'diameter', which is the minimum number of links or 'steps' in which some portion of all connected pairs of nodes can reach each other, revealed that the effective diameter is small for large 'real-world graphs' like an online social network; 3) the 'clustering coefficient' of nodes, which specifies the "ratio of the number of existing links over the number of possible links between its neighbours" (6), revealed that for any node within a tightly interconnected network the clustering coefficient is 1, meaning the relationships between a node and its neighbouring users is closely connected; and 4) finally the 'eigenvector centrality' property was analysed, in which a user with few connections could have a very high eigenvector centrality if those few connections were themselves well connected, and it was confirmed that this was the case within the social network user data analysed.

Alim et al [9] conducted a social network site analysis with the aim of discovering and highlighting vulnerabilities for users' personal data in SNSs. The investigators used an automated data extraction technique that consisted of web crawling with which they extracted personal data attributes and a list of top friends from MySpace profiles. Their methodology for data collection and analysis included the following key steps: 1) data pre-processing, which involved the analysis of the HTML structure of a profile, this step is needed in order to help in the design of the tables in the repository into which the collected personal data and list of top friends will be inserted for analysis; 2) specify the URL of profile, where the URL of the social network profile is used as a parameter to facilitate the extraction of personal data; 3) visit the specified profile webpage; 4) extract personal details from the profile; 5) extract a list of friends and their profile addresses and insert them into the repository; 6) the structural features of the repository are then automatically analysed using the 'Breadth-First-Search (BFS)' graph traversal algorithm, and a graph automatically generated to show vulnerabilities between "Friends" represented as nodes on the graph.

The results of the data collection and analysis conducted by Alim and colleagues found that, in regards to the extraction process, one problem in extracting data from the MySpace social network was that profiles differed depending on the type of profile and the users' preferences, making it problematic to implement the data extracting code. On the other hand, past analyses of the structure of various social network sites revealed that they employed a standard format, so even though many of the profiles were private, personal data such as gender, age, nickname, and the location of users, could still be extracted. This result indicated that users are particularly vulnerable to "social engineering attacks", or profiling, through inferences made from the personal details disclosed in the users' profile [9, p. 202].

Regarding the processing and analysis stage, a social network graph was generated based on the collected personal details and friends lists of all profiles crawled. This was done in order to analyse the graph for characteristics that could influence user's vulnerability through the spread of personal details in social network profiles. The graph generated was a "directed multigraph" since it was important for the study to analyse the flow of information in order for the direction of profile links or relationships – i.e. between a users' "Top Friends" and "Friends" represented by 'nodes' on the graph – could be known [9, p. 203]. The analysis for vulnerability provided by the graph indicated that profiles with a high number of "Top Friends" could be in danger from the possibility that of many of the users' so-called "Top Friends" were in fact strangers. This can lead to increased vulnerability in regards to the users trusting more and more people they do not know with access to their personal details, especially since these personal details are the kind that can be used to crack authentication questions or system identifiers, for example, "a pets name" [9, p. 206].

Thelwall et al [10] conducted a "sentiment analysis" of the social network site MySpace by crawling the web content – i.e. text, comments, html code, images videos – of user profiles in order to answer two related research questions: "How common are positive and negative emotions in social network comments?" and "Are there gender and age differences in the extent to which emotions are expressed in public MySpace comments?" [p. 194]. Their data collection process involved gathering a large random sample of MySpace comments. Because MySpace comments can contain images, video, irregular fonts, etc, these components were removed in order to retain only the textual data. Profiles of 30,000 MySpace members were randomly downloaded using an automatic crawling sampler. This dataset was reused from a dataset collected in a previous study [10]; however, the analysis in the current study was unique. The aim of the study by Thelwall [10] and colleagues, unlike the studies conducted by Alim et al [9] and Catanese et al [8], outlined above, was not to reveal vulnerability to the privacy of SNSs user's personal data, but rather to discover the correlation between positive and negative comments, found in the SNSs web content such as text, and gender. Their results found that positive sentiment was present in two thirds of the U.S. MySpace profiles and that women were the primary senders and receivers of positive sentiment within the social network site.

The above methodologies for social network data analysis reveal several common steps in the data collection and analysis process: data preparation; collection of raw data until extraction process (e.g. crawler) concludes its cyclical sampling; cleaning of data to make sure there is no duplication of information; and analysis of data gathered, usually represented in graph form, with 'nodes' of a graph representing users and the structure of the graph representing the relationships between users.

4. Demonstrating the Threats of Data Mining to the Privacy of the Users of Facebook and/or Other SNSs

4.1. The Scenario

Table 1. A sample dataset of Facebook users

Age	Nationality	Own Pic/week	Exposure	Status/ week	Ratio
21	Saudi	7	High	3	0.1
19	Pakistan	2	High	12	5
33	Yemen	7	High	2	4
5	Saudi	2	High	7	0.1
19	Pakistan	2	Medium	12	3
25	Yemen	3	Medium	3	4
22	Saudi	7	Medium	7	0.1
33	Pakistan	8	Medium	8	5
21	Saudi	7	Medium	22	5
27	Saudi	8	Medium	33	1
22	Saudi	2	Low	22	7
26	Saudi	8	Low	2	2
21	Saudi	7	High	3	0.1
19	Pakistan	2	High	12	5
33	Yemen	7	High	2	4
35	Saudi	2	High	7	0.1
19	Pakistan	2	Medium	12	3
25	Yemen	3	Medium	3	4
22	Saudi	7	Medium	7	0.1
33	Pakistan	8	Medium	8	5
21	Saudi	7	Medium	22	5
27	Saudi	8	Medium	33	1
22	Saudi	2	Low	22	7
26	Saudi	8	Low	2	2

Data Mining can be used for extracting various sensitive patterns of the user in Facebook or other SNSs. The obtained patterns can then be applied on a user in order to discover some sensitive information of the user that he/she does not reveal, resulting in huge user discomfort and serious breach of individual privacy. For example, a malicious data miner can study the Facebook activities and profile information of his/her female friends belonging to a country having an extremely conservative cultural structure say Saudi Arabia. Based on the observation the data miner can then prepare a dataset (i.e. a two dimensional table where rows represent the records and columns represent the attributes) having information on the users. In the dataset each record represents a Facebook user and each column/attribute represents a property of the user. Examples of the attributes include "Ratio of the Number of Opposite Sex Friends to the Same Sex Friends" (Col 6 in Table 1), "Number of Own Picture Uploads Per Week" (Col 3 in Table 1), "Level of Exposure of the Pictures (i.e. how exposed the person is in the pictures)" (Col 4 in Table 1), and "Number of Status Changes in Facebook per week" (Col 5 in Table 1).

We now give a hypothetical dataset, as shown in Table 1, created from Facebook activities and supplementary knowledge of a malicious data miner.

A data miner can apply an existing clustering algorithm on the dataset in order to cluster the records (users) where similar records are grouped together in one cluster and dissimilar records are grouped together in different clusters. In order to simulate the actions of a malicious data miner we implement an existing Fuzzy c-means clustering technique [11], which has been shown in the literature to be better than many other existing techniques, and then apply the technique on our example dataset (see Table 1) and thereby produce three clusters as shown in Table 2.

4.2. The Data Mining Analysis of a Malicious Data Miner

The data miner can then study the properties of the records belonging to a cluster. For example, a common property of Cluster 1 is having high number of opposite sex friends compared to the number of same sex friends. Some other common properties are having reasonably high number of Status updates per week, and low number of own picture uploading per week.

However, some common properties of Cluster 3 are quite opposite where the users have generally low ratio of opposite sex friends to same sex friends, low number of status updates per week, but high number of own picture uploading per week where the uploaded pictures have high exposure.

Such an observation may make the data miner more curious about the friends falling in Cluster 3 resulting in more careful analysis and follow up observations on the activities (both online and offline) of the friends. Therefore, falling in a group with such patterns can cause an uncomfortable situation for a female in a conservative society like the Saudi society where such a breach of private information of a female member of a family can cause serious damage to the reputation of the family.

Moreover, since the dataset has been created based on the Facebook information of the friends (female) of the data miner, he/she is likely to have supplementary knowledge on the users. By a careful analysis (based on his/her supplementary knowledge) of the friends falling in Cluster 3 the data miner may come to a conclusion that these friends are generally less conservative and "Willing to date". At this stage the data miner may reasonably come to a conclusion that any female Facebook user having similar properties is possibly also "Willing to date". In data mining, this is generally known as labelling the records with class values.

Based on this understanding, any unknown Facebook users who have similar properties as the properties of Cluster 3 can be categorised as "Willing to date" (in data mining, this is often called classification) and therefore approached inappropriately by the malicious data miner resulting in huge social discomfort for the user.

Without the use of data mining techniques and systematic analysis of unprotected Facebook information it would not be a trivial job to discover such a pattern especially from a big dataset having a huge number of Facebook users and huge number of possible attributes. Note that in this article we have only used a limited number of records in the sample dataset just to facilitate an example for discussion.

Table 2. Results of the application of clustering on the sample dataset of Facebook users

Cluster	Age	Nationality	Own Pic/ week	Exposure	Status/ week	Ratio
Cluster 1	19	Pakistan	2	High	12	5
	19	Pakistan	2	Medium	12	3
	25	Yemen	3	Medium	3	4
	22	Saudi	2	Low	22	7
	19	Pakistan	2	High	12	5
	19	Pakistan	2	Medium	12	3
	25	Yemen	3	Medium	3	4
	22	Saudi	2	Low	22	7
Cluster 2	22	Saudi	7	Medium	7	0.1
	33	Pakistan	8	Medium	8	5
	21	Saudi	7	Medium	22	5
	27	Saudi	8	Medium	33	1
	26	Saudi	8	Low	2	2
	22	Saudi	7	Medium	7	0.1
	33	Pakistan	8	Medium	8	5
	21	Saudi	7	Medium	22	5
	27	Saudi	8	Medium	33	1
	26	Saudi	8	Low	2	2
Cluster 3	21	Saudi	7	High	3	0.1
	33	Yemen	7	High	2	4
	35	Saudi	2	High	7	0.1
	21	Saudi	7	High	3	0.1
	33	Yemen	7	High	2	4
	35	Saudi	2	High	7	0.1

5. Concluding Remarks

This article explored the potential of data mining as a technique that could be used by malicious data miners to threaten the privacy of SNSs users. Using a hypothetical dataset, an existing clustering algorithm [11] was applied on this dataset to demonstrate the ease at which characteristics about the SNSs users can be discovered and used in a way that could invade their privacy.

The application of the clustering algorithm [11] on the above hypothetical dataset has shown that the threats from data mining on individuals' privacy can be serious. A malicious data miner can cluster his/her friends into different groups. He can then analyze the properties of the friends belonging to a group. Additionally, from his supplementary knowledge on the friends he can also add a label on them. Based on the extracted properties of the friends and the label he/she can next classify other Facebook users (who are not even in his/her friend list) into the label resulting in a breach of privacy. In this study we show an example where the data miner can classify other users into "Willing to Date" due the similarity of their properties and the properties of a group of his friends. In this example, the label "Willing to Date" is his supplementary knowledge on a group of his friends. Note that this is just an example of a possible attack from a malicious data miner on the individual privacy of the Facebook users. There are many other possible ways to attack the privacy of the Facebook users.

In this study we suggest that, to preserve privacy online, it is important that users mask their data or hide some information such as date of birth, address and other identifying information. This way, even if an unknown malicious data miner can classify a user, for example, as 'Willing to date', it can be difficult for the miner to locate the user and thus harass her. At the same time, if the properties of a cluster are known to the users they can deliberately design their activities in a way so that she is not classified as 'Willing to date'. For example, if a user (originally belonging to Cluster 3) has Exposure = "High" then she can deliberately change it to "Low" in order to avoid being grouped with other users in Cluster 3.

In sum, if all (or most) SNSs users place some false and misleading information about themselves (as we are suggesting), then a data miner will face difficulties to conclude with certainty whether or not the SNSs users really have a connection with a particular cluster.

REFERENCES

[1] D.M. Boyd, N.B. Ellison. Social network sites: Definition, history, and scholarship, Journal of Computer-Mediated Communication, 13: 210-230, 2007.

[2] R.K. Sar, Y. Al-Saggaf. Social networking sites' tracking of unintentionally shared information. First Monday, 18(6), 3 June 2013, 2013.

[3] H.T. Tavani. Ethics And Technology: Controversies,

Questions, And Strategies For Ethical Computing (3rd ed.). John Wiley: Hoboken, N.J., 2011.

[4] H.T. Tavani. Informational privacy, data mining, and the Internet, Ethics and Information Technology, 1: 137-145, 1999.

[5] I. Ting. Web mining techniques for on-line social networks analysis, International Conference on Service Systems and Service Management, June 30 2008-July 2: 1-5, 2008.

[6] R. Kosala, H. Blockeel. Web mining research: A survey, SIGKDD Explorations, 2: 1-15, 2000.

[7] L. Van Wel, L. Royakkers. Ethical issues in web data mining, Ethics and Information Technology, 6: 129-140, 2004.

[8] S.A. Catanese, D.E. Meo, E. Ferrara, G. Fiumara, A. Provetti. Crawling Facebook for social network analysis purposes, Proceedings of the International Conference on Web Intelligence, Mining and Semantics, May 25-27, 2011.

[9] S. Alim, R. Abdulrahman, D. Neagu, M. Ridley. Online social network profile data extraction for vulnerability analysis, International Journal of Internet Technology and Secured Transactions, 3: 194-209. 2011.

[10] M. Thelwall, D. Wilkinson, S. Uppal. Data mining emotion in social network communication: gender differences in MySpace, Journal Of The American Society For Information Science And Technology, 61:190–199, 2010.

[11] M. Lee, W. Pedrycz. The fuzzy c-means algorithm with fuzzy P-mode prototypes for clustering objects having mixed features. Journal of Fuzzy Sets and Systems, 160 (24): 3590-3600, 2009