Profile Similarity Technique for Detection of Duplicate Profiles in Online Social Network

Dr. M. Nandhini[1] , Bikram Bikash Das[2]

[1]*Assistant Professor, Dept of computer Science*
Pondicherry University, India
[2]*Department of computer science, Pondicherry University*
Kalapet, R V Nagar, India

***Abstract*— In the current generation social network has become a popular way to communicate with each other which are spread across diverse location around the world. In social network any user can find other users and make friendship and they can make friend circle online around the world and thus users can form their own network. An Individual user can have multiple social network accounts to keep in touch with friends in many social networking sites. Online Social Network users are not aware of the various security attacks like privacy violation, identity theft etc. Any user can create fake profiles with the name of real user. Other online social users will think it as real users and they might be responded to them which are not actually the real user. It makes the whole network quite confusing and frustrating. In this paper, we will provide a similarity technique which can analyze social network data based on attributes similarity. The proposed system can detect as many similar social network profiles as possible and analyse them in order to find whether it belongs to same or different persons. It makes other user easy to communicate with each other in a safe and efficient manner.**

Keywords—Social Network Analysis, Social Engineering Attack; Duplicate profiles; Global profile Database, Profile attributes matching, Suspicious profiles

I. Introduction

Social network makes our digital life become simple to have social relationship with friends online and social network websites such as Facebook, MySpace, Google+ for connecting peoples, YouTube for video sharing, Google+, LinkedIn for professional identity, Tweeter for updating daily tweets (messages) of some event etc. are becoming popular social network website used among peoples of all ages especially among youths. These social network sites are famous among internet users and users are interconnected to each other via online social relationship known as friendship. Various social network sites has been developed to gain their attraction among people where any use can get can membership by fill up a simple registration form. An individual user can make multiple accounts with his same attributes such as E mail id or mobile number in many social networks. [1]

An existing user can have many numbers of social network profiles within the same or different network. It makes social network vulnerable to attack by using someone similar attributes. Most of the cyber crimes are happening in social network sites. [2] Any user can make profiles with others attributes such a same name, college, age, profile images etc. Fake profiles are being created in all the social sites and victim personal information is becoming more and more vulnerable to attack by the attacker in various ways. In many cases Name can be same for many user but the other attributes such as profile image, qualification, address and mobile numbers all can't be same for multiple social network users[1][2]. The main idea behind the proposed approach is to find out as many social network users which have similar attributes and to find out the originality or real users from it. In recent Research reveals that almost 80% of profiles in face book are fake one. Any user can make some fake accounts by using others attributes to fool other users. [3]

Many users' like to disclose their personal information like phone no., date of birth, address etc in their profiles. Availability and revealing of such personal information might be the sources of profile data that the attacker is trying to get access to create similar profiles. Any other user can create Fake profiles in the name of the real user with that personal information and try to launch various attacks such as sending and posting irreverent messages and tries to fool others to get the confidential information.[4]

The proposed approach can detect as many similar social network profiles as possible based on the similarity of profile attributes and analyse them in order to find whether it belongs to same or different persons. The publicly available profiles and their attributes are extracted and then store them in global database of profiles so that their existence can be checked in many social networks. It helps the others user to have online social communication with each other in a safe and efficient manner.

II. Related Works

Many of cybercrimes are because of the facilities to create unlimited numbers of profiles by the same person within the same or different networks who try to act as real users and violating the rules. It makes other users difficult to identify who are the real users and who are fake. Thus social network analysis comes to the scenario which is a new area of research that involves analysing the network structure for the benefit of society so that online user can have social connection and conversation in a safe way with other user around the globe.

In recent scenarios it becomes a serious problem and many researchers have begun their research in identifying the real user's identity. Sophia Alim, Daniel Neagu ,Ruqayya

Abdurrahman and Mick Ridley (2011) had presented an approach for automated extraction of social network user data since the large amount of data in a web database are hidden which are not generally indexed by the search engines. They proposed a generalized method of finding fake or false profile identification. First a novel approach for extraction of personal data was implemented to extract profile details and a list of top friends from social network profiles and store in a data repository. An online social network graph will be generated from the repository data where the nodes represent peoples' profiles or group and edge represents the social connection or friendship relationship between them. [3]

They have presented an approach for detection of fake user profile by using an OSN graph which is generated automatically. Thus we can consider social network as a graph which is composed of vertices and edges as shown in the figure below. The vertices represent the individual user profiles and vertices represent the social relationship or friendship among them. Breadth First Search was used to travel across the network. [3][4]

Breadth First Search has been implemented to search the user profiles. The following figure shows that profiles 1, 3 and 5 are common to profile 2 and they are mutual friends to each other. Likewise others profiles relation can also be represented. The various structural features and their relation between the profiles of the graph have been analyzed to see how they contribute towards the vulnerability of a node.

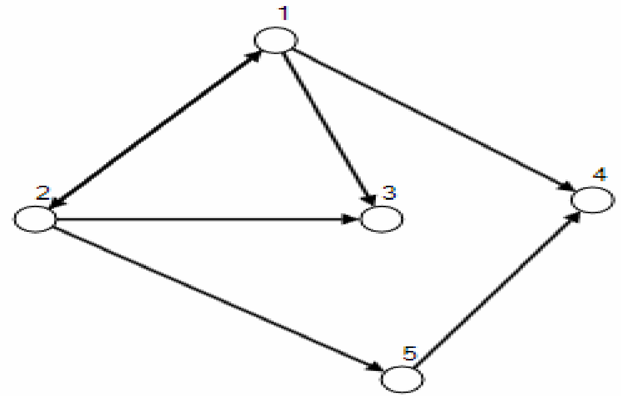

Fig : Graphical structure of social network data

The arrows in the above figure represent the 'is a friend of' relationship. The node presented in the figure is user's profiles and the arrows represent their social relationship such as friendship relationship among them.eg. 3 is a friend of node 1 and node 2 and so on.

Using Breadth First Search, Look at profile 1's friends network and analyze their relationship with other users and check to see whether they have been already exist there in the repository. So the friends of profile 1 which are node 2, 3 and 4 are checked to see if they are already exist or not.

If those friends are exist in the repository then their profile attributes along with a list of top friends are added to the queue. Check the next profile which is present at the front side of the queue. Repeat these steps until all profile has been extracted. [3]

Himanshu Gupta, A Arokiaraj Jovith (2013) had presented a Trusted Profile Identification and Validation mechanism where they have mentioned the following social attack and their influences in our day to day life. [5]

- Attackers may duplicate a victim's user online identity to launch various social engineering attacks.
- Attackers try to create fake profile with the intent to fool other social users with the name of some real users using their personal data.
- Attackers may spread false and sensitive messages to create panic situation in the public.
- Attackers may ask any social user to pay for some offer and services which are actually not exist in real case.
- Attackers may use the e-mail or mobile numbers of some existing user and they may be the victim of various social engineering attacks like identity clone attack, hacking someone's profiles to get confidential information, DOS attack etc.

To overcome the above security issues the proposed approach has been presented in order to find the real user identity to check which one is fake and which one is real user profile.

III. Limitations Of The Existing System

A. Protection of users privacy limits data collections

Even though social media seems to be a very open space, each and every social network has their own privacy settings where Users can have a certain level of privacy. E.g. Face book provide a facility to have users their own privacy settings where they can show limited number of information and hide their personal details without verifying them. So the genuine profile detail is not always possible. [3]

B. Social network profiles are prone to various attacks:

The users in the social network are related to each others in the form of friendship. Social users are prone to various attack such as identity clone attack where the intruders create some fake profiles with the name of some existing users and he can create the profiles with exactly same as some existing user that looks similar to the real user and the attacker might try to lunch various attacks.[7]

C. Social networks are more complex to analysis:

The users in the social network are related to each others in the form of social relationship and they can be represented in graphical format and these networks are complex to analyze.

D. Finding the real users identity is difficult to detect:

A user can create any number of profiles within the same or other social network with false identities to fool other users. Thus the whole social network become more complex and confuses other users with multiple identities. [5]

IV. OUR PROPOSED APPROACH:

Going by literature review, we have found out that the social network is having complex structure and to analyze their relationship and patterns is complex in nature. Moreover the existing processes require more computational time and involve many complexities.

In Social network very large amounts of personal information are being shared and posted online daily. Thus an anonymous user can retrieve the personal details of individuals and fake or false profiles which seem to be like the real users. Now a day's social users are more vulnerable to numerous social engineering attacks like Identity Clone attack, fake profiles creation, hacking etc. because of their personal attributes are easily available. Hackers are always trying to find loopholes in the existing system.

Online social network allows its users to create infinite numbers of profiles to connect with social relationship with others. A user can create many numbers of profiles within the same or different social network with different identities. The social security is also a major problems associated with this. A user can make as many profiles as he want. Some other user can create profiles with the same name of already existing users with the intent to fool others user and to get the personal information. It became difficult to detect which one is real and which is fake one.

The proposed system to be developed is an application through which the user to have the facility to have a Web-based user profiles search mechanism. This facility could be used to search for individual user in a number of social networks and produce a consolidated output along with summary of with duplicated user profiles.

First we created our own social network profiles to investigate different possible structures and attributes. Another application might be a "meta" social network website which has a single environment for user through which they can search and access other profiles details. Users could connect their accounts with other social networks and the Meta social network website would consolidate all their information and friends' networks. It provides the user to have a simple and effective way to communicate and to keep up-to date with their friends' activities across all the social networks from a single environment.

A social network is a place where we trust each and every user based on their online identities only. But most of the people out there are not real account but the fake people with false identities who are trying to do some malicious activity known as a scammer. A scammer can represent himself as a real user to get the personal details of other users. Even he can chat online with other users, build trust among potential users by posting unique and original ideas with each other, steal money belongings or even life because cyber criminals are targeting social networking sites to steal money.

The main purpose of the proposed work is to provide a mechanism to solve these issues with the help of using global database of profiles where the attributes of various social network profiles are stored.

The proposed tool can detect duplicate profiles exist in social domains, and conduct a case study. First is the data collection phase where we collect some publicly available user profile data sets that has to be extracted and place those details information into a database called global database in order to find the duplicate profiles. Then we analyse and study the pattern of the publicly available profile information to identify the real user's identity from a large datasets.

V. PROPOSED DESIGN APPROACH

In this section we have outlined the design approach of the proposed work in diagrammatic form for identifying duplicate profiles and checking the existence of similar profiles in many social networks and validating the genuine profiles.

The proposed model comprises of three components and the following section describes it one by one. In the following diagram the proposed work has been presented where the whole work flow is divided into 3 processes.

- Profile identification process.
- Profile evaluation to check the existence in Social networks
- Profile verification process to check whether the profile belongs to the same user of some different user profiles.

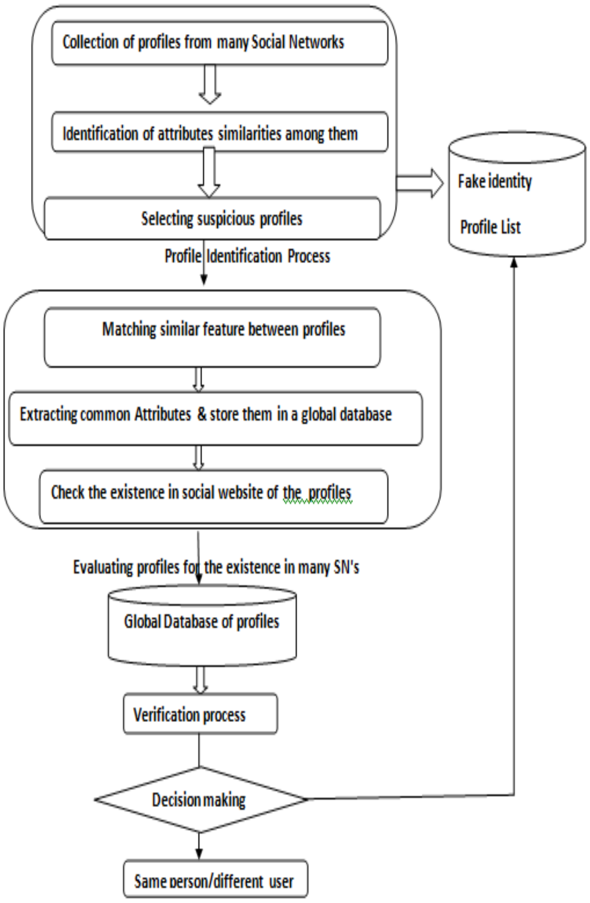
Collection of profiles from many Social Networks
Identification of attributes similarities among them
Selecting suspicious profiles
Profile Identification Process
Fake identity
Profile List
Matching similar feature between profiles
Extracting common Attributes & store them in a global database
Check the existence in social website of the profiles
Evaluating profiles for the existence in many SN's
Global Database of profiles
Verification process
Decision making
Same person/different user

Fig: Proposed Fake Profile Detection Process Diagram

The main objective of the proposed approach is to find the real users identity across various social networks with the help of using global database of profiles where the attributes of various social network profiles are stored. An algorithm that detects similar profiles in various social networks and extracting their profile attributes from various suspicious profiles which helps to identify the real users among them and find any duplicate profile existence within the same or different social networks by using their profiles similarity among them.

The proposed profile similarity technique in social networks helps to identify a particular online user who has multiple social networking accounts within the same or many different social network sites and map and compare his/her profile's attribute values with other similar online user in the same or different social network to do online search easier and to improve the internet search by using the global Database of profiles.

The primary advantage of the proposed approach:

1. Since all the profiles details are stored in the global database instead of storing them into different database of profiles.
2. Here the common profile attributes of users which are common in many social network are extracted and stores those attributes in the global database so that the proposed mechanism can easily find the duplicate profiles and to make search process faster and check the existence of same profiles within the same or many social networks.
3. Instead of checking duplicate profile existence in different social network individually, the global database of profiles is used where all profiles details are stored.
4. Memory consumption can also be reduced and the internet search can also be improved by utilizing social network / user generated content to improve search.
5. To stop and prevent the rapidly increasing fake profile creation around the world.

VI. PROPOSED DESIGN APPROACH DESCRIPTION:

There are increasing numbers of social network users day by day. This increase of social network users make the whole network vulnerable to attack since there is no restriction of creating profiles and anyone can create accounts with the same name of others profiles with the intent to fool others or to post some irreverent personal information of some already existing users without the intention of the real users.

a. Profile identification process:

Step1. Collection of profiles from many social networks: This is the information gathering step from different social networks.

Step2. Identifying similar attributes among them: To make a relationship between two or more person across social networks.

Step3: selecting only the suspicious profiles:

b. Evaluating profiles to validate the existence of user profiles in many Social Networks:

Step 1. Check and Matching similar attributes among profiles (matching profiles fields): Here the two profiles for their similarity based on their HTML structures of profiles are checked and put them in a common database of profiles. It results in the following two outcomes:

Exact matching: The first category of matching analyze the user profiles by using some matching function such as string comparison to check whether there exist any two data fields which are exactly similar. Matching functions of this type produce a Boolean result. E.g. the exact field matching function to match attributes such as "usernames".

Partial matching: The second category of matching functions analyzes the user profiles which match the parts of related profile attributes. They are more useful in cases where the user profile data has many redundant values such as many abbreviations, misspellings or some missing values (e.g., address). Some function might be used to enable the matching of similar data values which are partly related to each other. An example of such a function is address or location matching.

Step2: Extracting the common attributes based on similarity scores:

Step3: check the existence of social sites of the profiles (e.g. face book, tweeter, YouTube etc.)

c. Profile Verification to identify the real social User:

Whenever a user search and type for any user profile, many profiles show up with the similar name out of which one will be the real user that the intended user is searching for. But users cannot guarantee that whether profile we are searching is fake or real. The current existing approach verifies those profiles information by manual process and checks for whether the name is a well known person or it checks to match with other similar user. If that profile is some known person likes celebrities or any politician then it will check whether the profile is connected with any official sites or govt. approved pages or if they are connected to any TV shows, Interview etc. If the profile seems to be doubtful then they will be asking proof like a faxed ID. The main drawback with this current approach is that it is very manual, takes lot of time to process it and so overwhelming. Verification methods are more generally used to find the original profiles and to authenticate of the real user identities in social networks. It helps the other users to get the genuine, real users and trustworthy information and discover high-quality sources of information and to maintain trust that the legitimate sources of information are genuine.

The proposed process verifies the users in the following ways:

Step1. Checking the friends network (Mutual friends relationship)

Step2. Maintaining trust or distrust among social networks users.

Step3. Verification using questionnaire and validate of answers.

Step4. Ask users to upload any government ID proof like PAN card, adhaar card copy etc.

Step5. Verifying those Govt. ID card information using matching techniques with the user profile attributes

Step6. Decision making to get the final result (same person or different user).

VII. EXPERIMENTS

In this following section, the experimental methods and techniques are presented for the proposed approach to validate the fake profile detection process. As face book is the most popular social network. Therefore dataset of publicly available Face book user profiles information is extracted and the various users' attributes like Profile Name, Address, Interests, User Home Page, Likes and Friend relationship are added to the global database. Then we relate those attributes to other social network to check whether there is any other similar user profile exists. After that the real user detection process has been evaluated to validate them.

The detection framework cannot be implemented into real system as such activity violates the terms and conditions with OSN sites. Here a set of profile attributes and their online identities is considered as fake profile identities and their relationship are shown graphically as node and vertices where the node represents user profiles and the edge between then represents their friendship relationship as shown in the figure below.

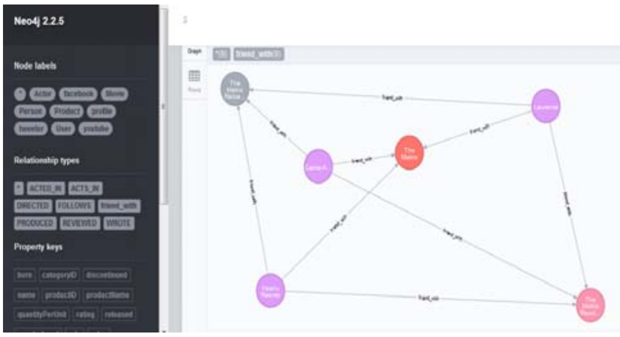

Fig: Graphical structure of Social Network user profiles

Software analysis:

For the proposed approach, the Neo4j 2.2.5 tool is used for analysing the profiles dataset in graphical format as shown in the figure above. The Proposed detection approach is first implemented on publicly available datasets. Software like PHP (Hypertext pre Processor) is used for describing the HTML structure of profiles in the web and store those detail attributes of profiles in My SQL database which is termed as global database. Then compare their attributes of profiles and extract those similar user profile data to represent them in social graph and after that it will validated using the proposed detection model.

VIII. FUTURE PLAN:

Social network analysis deals with complex network structure of profiles and which are rapidly changing over time. This propose work analyze those publicly available social network data and present an approach for finding duplicate profiles by profile similarity technique and store them in the global database and provide mechanism to check the real user identity.

Biometric authentication can also be implemented to validate the real user identities based on finger prints, signatures, facial structure, person's voice etc. which is unique for each user.

This work done can be improved by Future research by implementing profiles matching algorithm to verify suspicious profiles and the user verification process can be improved by using biometric authentication system to validate the suspicious profile. Data mining techniques can also be used to improve the duplicate profiles detection process and to analyze them to see in how many social networks those similar profiles are existed.

IX. CONCLUSION:

Our approach identifies the real user profile in social networks.This proposed approach first identify the common attributes to check in the similar profile is exist in SN and to check the real user.Then the profile matching mechanism has been implemented by comparing the similarity between the attributes and provide a decision algorithm that justify our approach. Thus it solves the problem of creating fake profile and detects the genuine users.

REFERENCES

[1] G. Kontaxis, I. Polakis, S. Ioannidis and E. Markatos, " Detecting Social Network Profile Cloning" In Proceedings of IEEE International Conference on Pervasive Computing and Communications, pp. 295-300, 2011.

[2] L. A. Cutillo, R. Molva and T. Strufe, "Safebook: A Privacy-Preserving Online Social Network Leveraging on Real- Life Trust", IEEE Communication Magazine, pp. 94-101, 2013.

[3] Sophia Alim, Ruqayya Abdulrahman,Daniel Neagu and Mick Ridley, "Online social network profile data extraction for vulnerability analysis", International journal of Internet Technology and Secured Transactions, Vol. 3, No. 2, pp.197-200, 2011.

[4] G. Kontaxis, I. Polakis, S. Ioannidis and E. Markatos, " Detecting Social Network Profile Cloning" In Proceedings of IEEE International Conference on Pervasive Computing and Communications, pp. 295-300, 2011.

[5] Himanshu Gupta, A Arokiaraj Jovith, " Trusted Profile Identification and Validation Model" International Journal of Engineering Research and Development e-ISSN: 2278-067X, p-ISSN: 2278-800X,Volume 7, Issue 1 (May 2013), PP. 01-05

[6] C. G. Akcora, B. Carminati and E.Ferrari, "Network and profile based measures for user similarities on social networks", In Proceedings of the
IEEE 11th International Conference on Information Reuse and Integration (IRI), pp. 292-298, 2011

[7] Fatemeh Salehi Rizi, Mohammad Reza Khayyambashi, and Morteza Yousefi Kharaji , "A New Approach for Finding Cloned Profiles in Online Social Networks", International Journal of Network Security, Vol. 6, April 2014.

[8] C. G. Akcora, B. Carminati and E.Ferrari, "Network and profile based measures for user similarities on social networks", In Proceedings of the 2012 IEEE 11th International Conference on Information Reuse and Integration (IRI), pp. 292-298, 2012.

[9] M.Conti, R.Poovendran, M.Secchiero, "FakeBook: Detecting Fake Profiles in On-line Social Networks", In IEEE/ACM International

Conference on Advances in Social Networks Analysis and Mining, 2012

[10] Acquisti (2005), 'Information Revelation and Privacy in Online Social Networks.' Paper presented at of the ACM workshop on Privacy in the electronic society, Alexandria, USA, p.142-250,November 7,2005

[11] L. Jin, H. Takabi and J. Joshi, " Towards Active Detection of Identity Clone Attacks on Online Social Networks", In Proceedings of the first ACM conference on Data and application security and privacy, pp. 27-38, 2011.

[12] G. Kontaxis, I. Polakis, S. Ioannidis and E. Markatos, " Towards active Detection of Social Network Profile Cloning Attack" In Proceedings of IEEE International Conference on Pervasive Computing and Communications, pp. 295-300, 2011.

[13] A. Mislove, B. Viswanath, K. P. Grimaldi, P. Druschel, "You Are Who You Know: Inferring User Profiles in Online Social Networks" In proceedings of the 3th ACM international conference on web search and data mining, pp. 251-260, 2010.

[14] E. Zheleva, L. Getoor, "Join or not join: the Illusion of privacy in social networks with mixed public and private user profiles", In proceedings of the 18th international conference on World wide web, pp. 531-540, 2010.

[15] Florian Kerschbaum and Andreas Schaad," Privacy-preserving social network analysis for criminal investigations", In Proceedings of the 7th ACM workshop on Privacy in the electronic society ACM, New York, USA, p.9-14, 2008.

[16] Leyla Bilge, Thorsten Strufe, Davide Balzarotti and Engin Kirda, "All your contacts are belong to us: automated identity theft attacks on social networks", In Proceedings of the 18th international conference on World Wide Web ACM, New York, USA, p.551-560,2009,

[17] Sanjiv Sharma and R. K. Gupta," Improved BSP Clustering Algorithm for Social Network Analysis" In International Journal of Grid and Distributed Computing Vol. 3, No. 3, September, 2010.

[18] Fatemeh Salehi Rizi, Mohammad Reza Khayyambashi, and Morteza Yousefi Kharaji, "A New Approach for Finding Cloned Profiles in Online Social Networks" International Journal. of Network Security, Vol. 6, April 2014.

[19] The Anh Dang, and Emmanuel Viennet, "Community Detection based on Structural and Attribute Similarities": The Sixth International Conference on Digital Society In ICDS, Vol. 2, No. 4, p.351 360, September,2012.