JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DETECTION OF CLONE PROFILES IN FACEBOOK USING SIMILARITY MEASURES AND CLASSIFICATION ALGORITHM

Sowmya P
Assistant Professor
Department of Computer Science and Engineering
Nitte Meenakshi Institute of Technology, Bengaluru, India

Abstract: Online Social Network (OSN) is a network of users with similar areas of interests. With the gaining popularity of OSN day by day, privacy and authentication processes related to it are compromised by attackers. Users of social networks are experiencing extremely risky security breaches. One of the most dangerous cyberthreats is profile cloning, in which data from already-existing users is taken to generate duplicate profiles that are then used to compromise the identity of real users. They then cause other attacks like phishing, stalking, spamming etc. using these clone profiles. So, detection of these types of duplicate profiles in online social sites is essential to avoid adverse effects caused by them. This paper proposes a detection algorithm that can identify Facebook clone profiles. Two techniques are applied to the detection of profile cloning, one utilizing similarity measures and the other employing the C4.5 decision tree technique. Two sorts of similarities are considered in similarity measures: similarity of attributes and similarity of network interactions. By creating a decision tree and taking information gathered into account, C4.5 finds clones. Finally, both the methods are compared to examine how well these two approaches recognize clone profiles.

Index Terms- Clone, Identity Theft, C4.5, Attribute Similarity, Network Similarity

I. INTRODUCTION

Most of the world's population uses well-known social networks like Facebook, Twitter, LinkedIn, Instagram, etc. to exchange information. A new era of networking has emerged because of the social networks' accessibility. OSN members post a lot of information on the network, such as images, videos, education, location, school or college names, phone numbers, email addresses, residential addresses, family relationships, banking information, and employment information. Severe repercussions might result if this private information falls into the wrong hands. Because the majority of OSN users are ignorant of the security risks that social networks provide, they are easily targeted by these assaults. If children are the intended victims, the risk is quite high.

Identity theft of this kind takes the form of profile cloning, in which genuine users' credentials are taken to produce clone profiles that may be used in a variety of assaults. Attacks using profile cloning may be divided into two categories: cross-site and same-site attacks. Same Site Profile Cloning refers to the process of creating duplicate profiles on the same website where the original user already has an account. Cross Site Profile Cloning is the process of using user credentials from one website to establish a duplicate profile on another website where the user does not have an account. Famous people's clone profiles are made and then exploited inappropriately. Clones are typically created to malign them.

The signup procedure for social networks is straightforward to attract more users. These are being used by the attackers to steal information. As a result, the number of duplicate and false profiles is rising alarmingly.

II. LITERATURE SURVEY

Clone or duplicate profiles are now a very severe concern to social network members. Attackers construct these profiles, which they then abuse in a variety of ways. Therefore, a clone detection technique is crucial to catching these con artists who use people's trust to collect personal data and build replica profiles. Numerous scholars have studied this topic and have put up various strategies for recognizing these kinds of profiles in social networks. Below is a discussion of a few of these approaches.

A defensive strategy with two phases—the verification phase and the decision-making phase has been put forth by Gordhan Jethava and Udai Pratap Rao [2] to shield users from cloning assaults. The verification and decision-making steps are where the friend request goes once it is submitted. In the Verification step, similarities across profiles in terms of attributes, friend lists, and behaviors are measured. Important attributes are identified and used as a similarity measure to compare the profiles. The friend lists similarity uses friend lists to compare similarity between profiles. And behavioral similarity helps in measuring the strength of relationships between profiles and in authentication of users to the social network. In order to make a judgment, the Similarity Score (SS) is compared to predetermined thresholds. The friend request is regarded as coming from an actual user if SS is below a certain threshold. Friend requests from accounts that have been duplicated are not accepted.

S Revathi and Dr M Suriakala [3] have proposed a method for clone profile detection in Facebook using Network Theory. Node Similarity Communication (NSCM) algorithm is used for detecting clones based on malicious activities of users in social networks. The activities that are considered are- updates, posts, comments and photos. Comparing thresholds of attributes and similarity of networks, clones are detected. The process involves creating an account, performing user operation, monitoring, finding recent activities and detecting cloned profiles. The proposed methodology worked with an accuracy of 93.14% in detecting clones.

A prototype to determine if users have been the targets of cloning attempts has been proposed by Georgios Kontaxis, Iasonas Polakis, Sotiris Ioannidis, and Evangelos P Markatos [4]. Working with LinkedIn is the prototype. A user's profile is mined for information, and then a search is conducted in OSN to locate profiles that match the information taken from the user's profile. Based on the similarity of the attribute values, a similarity score is computed. The specific profile is considered a clone if the similarity score exceeds the threshold value. The drawback of this proposed method is that it looks for exact string matches between the profiles. So, it cannot detect wrongly typed data or purposefully injected mistakes by attackers.

Brodka, Mateusz Sobas and Henric Johnson [5] in their paper proposes two novel methods of profile cloning detection in Facebook. The first method is based on the similarity of attributes from both profiles and the second method is based on the similarity of relationship networks. Attribute Similarity based Profile Cloning Detection (ASPCD) algorithm is used to compare 5 attributes namely First Name, Last Name, Gender, Localization and Education. Gender and Localization are compared using a simple string-matching algorithm where the similarity index is set to 1 if both the profile values are same else it is set to 0. First Name and Last Name are compared using Dice coefficient. Education is compared in the form of Hashmap where each college has a unique id, which is simple to compare. Similarity of network relationships are calculated based on Network Similarity based Profile Cloning Detection (NSPCD) algorithm where mutual friends are considered for calculating similarity index. The drawback is that the proposed method detects clones based on attribute similarity and network similarity separately and both the similarities are not combined.

In their study, Kiruthiga S, Kola Sujatha P, and Kannan A [6] describe how the detection of clone attacks is based on user click patterns and the duration of user activity. The specifics for each user's information are classified in this case using the Naive Bayes Classifier. To group similar networks, K-Means clustering is employed. Utilizing Cosine similarity to uncover similarities and Clone Spotter to identify cloning on Facebook, performance may be improved. In the K-Means clustering algorithm, the characteristics Uid, Weight, Network1, Network2, and Network are regarded as data points. For each cluster, the mean value is computed. Every piece of information about the user, including First Name, Last Name, Age, Date of Birth, Sex, Number of Visiting Friends, Total Friends, Hometown, Current City, Users Click Pattern, and Users Action Time Period, is recorded by the server for each user friend request from X to Y. It is efficient to find clones using the Clone Spotter method. The cosine similarity metric may be used to compare two product vectors. For each user, the following data is reviewed to see whether there are any similarities: visiting friends, total friends, hometown, current city, action time pattern and clicking pattern.

From the research, it can be concluded that most approaches use straightforward string-matching algorithms as a similarity measure. As a result, it is unable to successfully detect clones since it cannot overcome incorrectly entered data or mistakes that are purposely introduced. Furthermore, most approaches are used with offline datasets and cannot be used in a real-world setting.

III. PROPOSED WORK

3.1 Clone detection using Similarity Measures Algorithm

A major social threat is now Profile Cloning. In social networks, private information such as a phone number, email address, name of a school or college, company, birth date, location, etc. is publicly accessible, making it easy for hackers to build false identities. These accounts are then utilized in a variety of cyberattacks, such as phishing, spamming, cyberbullying, etc. Even the well-known individual or the company may be defamed by them. To increase the security of users' social lives, a detection approach that can identify duplicate profiles on Facebook has been developed. The suggested system's architecture, as seen in Fig. 1, is used to identify clones using similarity metrics.

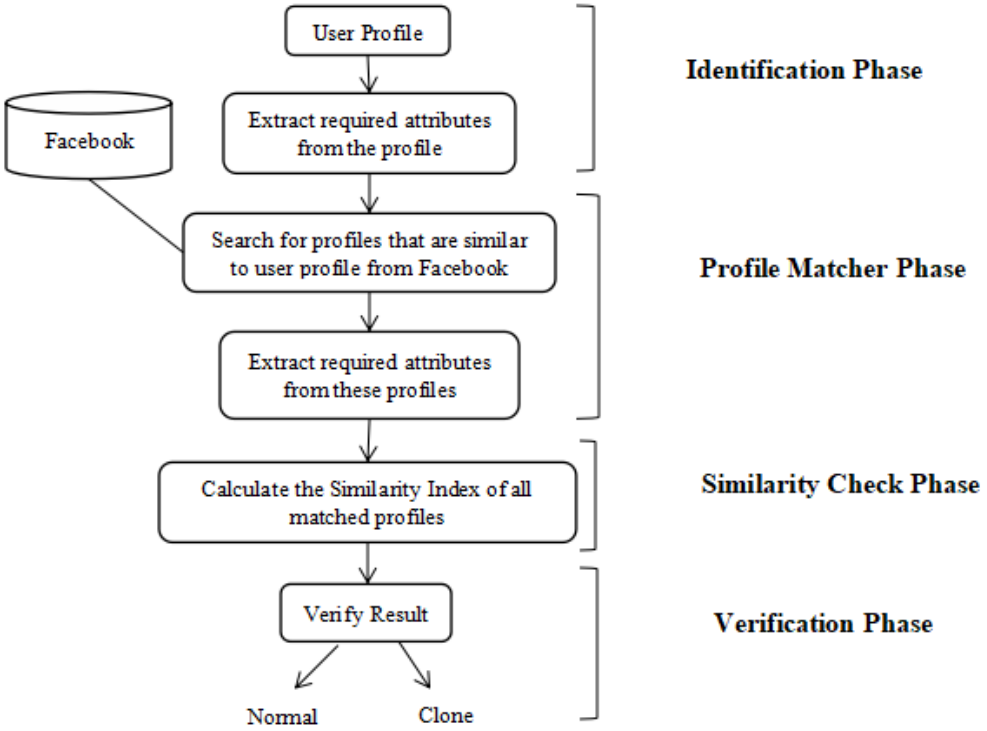

Fig 1: Architecture of the proposed system

The proposed system uses the following steps to detect cloned profiles

- Select User Profile
- Extract required attributes from the profile
- Find every profile that resembles the user's profile and take the necessary attributes out of it.
- Calculate the Similarity Index of all matched profiles and compare it with that of user profile
- The profile is referred to as a clone of the specified user profile if the Similarity Index exceeds the threshold. If not, it is a typical normal profile.

3.1.1 Identification Phase

This is the initial phase where the user who wants to check whether his/ her profile has been cloned is selected. This user profile is analyzed to extract the distinguishing attributes from the profile. The following attributes are extracted from profile using Graph API

First Name	Location
Last Name	Hometown
Gender	School Name
Birth Date	College Name
Email	Friends List
Work	

User Identifying Information is the name given to this data, which is then forwarded to the following stage. The Graph API is used to retrieve these attributes from Facebook.

Extraction of attributes from Facebook Profile

Facebook Graph API is utilized to retrieve the necessary attributes from Facebook. It is a programmatic API that uses HTTP to retrieve data from Facebook. The Graph API is made up of nodes, which are essentially "things" like a user, a photo, a page, or a comment edges - the connections between those "things" like a Page's Photos or a Photo's Comments fields and information about those "things" like a person's birthday or a Page's name.

Graph API Explorer: It is a basic tool used to search, add, and remove data from Facebook. An Access Token is necessary to query Facebook. The access token identifies a user app and approves the request for permissions. The various permissions that are available are first name, last name, email, phone number, birthday, hometown, location, gender, education, friend list etc. This data can be extracted from Facebook using an access token.

To get the data from Facebook, the Graph API must get permission from the user to get that data. When users log on to graph API using Facebook login, they receive a request to grant permission to the requested data. Users can grant or deny the permissions requested.

3.1.2 Profile Matcher Phase

Once the required attributes are extracted from the profile, similar profiles are searched in the network based on name and required attributes are extracted from them using graph API. And the friend list is extracted to compare the profiles for network relationships.

3.1.3 Similarity Check Phase

The similarity check phase is applied to the profiles that emerged from the prior phase's search procedure. To get the similarity index between the profiles, each of these profiles is compared with the user profile. Attribute similarity and Network similarity are the two forms of similarity that are employed. The similarity of the attributes between two profiles serves as the basis for measuring attribute similarity. To compare the network relationship similarity between two profiles, network similarity is based on the friends lists of the two profiles. If the aggregate similarity index of attributes and networks exceeds a predetermined threshold, the profile is questioned as a clone. Otherwise, it is a typical or authentic profile.

Attribute Similarity

Similarity of attributes between the profiles are considered for Attribute similarity. The attributes that are considered for similarity are First Name, Last Name, Gender, Birth Date, Email, Work, Hometown, Location, School Name and College Name. Here, some similarity measure algorithms are used to find similarity namely - Cosine similarity, Levenstein distance, n-gram similarity and exact string matching.

Cosine similarity - When two vectors are given, cosine similarity returns a value of 1 if the vectors have the same orientation, a value of 0 if the orientation is at 90^0, and a value of -1 if they have diametrically opposed orientations. The following is the cosine similarity formula to compare two vectors.

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\ \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Levenstein distance - It is applied to compare two sequences character by character for resemblance. By accounting for the quantity of insertions, deletions, or substitutions necessary to switch from one sequence to another, the Levenstein distance determines how similar two sequences are. The following formula may be used to calculate the Levenstein distance between two sequences with lengths of i and j.

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1, \\ lev_{a,b}(i-1,j-1) + 1_{a_i \neq b_j} \end{cases} & \text{otherwise} \end{cases}$$

n-gram similarity - By dividing the strings into unigrams, bigrams, trigrams, etc., n-gram similarity analyzes the similarities between two strings. The sequence of words should also be considered when comparing characteristics using the n-gram similarity method to determine how similar two strings are. The source of the n-gram commonality is

$$\text{n-gram similarity} = \frac{\text{The number of n-grams that both X and Y share}}{\text{N-grams with the most between X and Y}}$$

Exact string matching - It checks whether the two strings are the same. Values are set to 1 for similarity if they are equal and 0 for dissimilarity.

The total attribute similarity is determined by applying the following formula after computing the similarity of each attribute separately.

$$S_{att}(Pc, Pv) = \frac{\sum_{i=1}^{n} E_i(Pc, Pv)}{n}$$

where S_{att}- Attribute Similarity, Pc - Profile of clone, Pv - Profile of victim, n - Number of different attributes compared,
Ei (Pc, Pv) - Function returning the similarity of i^{th} attributes of Pc and Pv

Network Similarity

The calculation of network similarity uses mutual friend similarity between the profiles. To obtain the victim's sensitive information, the attacker consistently tries to make friends with the victim's friends. Whenever a friend request is received, without a second thought it is accepted by some Facebook users. This has been taken as an advantage by attackers where they create clone profiles and send friend requests and most of the time it is accepted without any cross verification and their sensitive information is easily exposed to attackers who create clones. So along with Attribute similarity, Network similarity is also considered to detect clone profiles efficiently. The formula to calculate Network similarity is as follows

$$S_{Net}(Pv, Pc) = \frac{(|MFF\ vc|)}{(\sqrt{|Fv|}.\ |Fc|)}$$

S_{Net} - Network Similarity Pv stands for victim profile, Pc for clone profile, MFFvc - Set of matching Pv and Pc mutual friends, Fv - Pv's group of friends and Fc - Pc's group of friends

The profile is handled as a clone if the NetSim value is higher than the threshold, else it is processed normally.

3.1.4 Verification Phase

The final stage is to manually check the results, which is done by the user. He is aware of which profile is his original profile and which is a clone. The threshold setting is a crucial element since it would be very challenging to manually check all the profiles for clones if there were too many alerts.

3.2 Clone detection using C4.5 Algorithm

To determine if the provided profile is a clone or not, this module uses the C4.5 algorithm. The categorization process uses the decision tree algorithm C4.5. With the information provided, this generates a decision tree. Each node in the tree has an attribute that separates sample sets into subgroups most efficiently. Entropy and information gain are the distribution factors employed in C4.5. The attribute with the greatest information gain determines which subtrees should be cursed again. The C4.5 method constructs a tree-like structure to detect similarities between attributes. This profile is being contrasted with profiles that are already stored in the database. Cloned profiles are those that match profiles in databases; otherwise, they are considered normal.

IV. EXPERIMENTS

Data extraction from Facebook profiles

Data from Facebook was extracted using the Facebook Graph API. The necessary information was taken from Facebook accounts of friends, coworkers and students who were willing to participate in this study with their consent. The Graph API facilitates data input into and extraction from Facebook's platform. Through programs, it also aids in data searching, posting new data, uploading photographs, etc. We require an Access Token to query Facebook. Graph API Explorer may be used to extract a variety of data, including First Name, Last Name, Gender, Birthdate, Hometown, Education and Work. After that, we may look for profiles that resemble the victim's profile using a query search engine. Finally, suspect profiles can be subjected to similarity measures and C4.5 algorithm to determine whether they are clones.

Evaluation Metrics

Based on the four main standard indicators, several assessment metrics are employed to assess the system's performance.

- True Positive (TP): Records that are successfully discovered using anticipated vectors are considered true positives.
- True Negative (TN): Records accurately recognized as Neutral recordings are true negatives.
- False Positive (FP): Records are those that the system mistakenly thought were present in one vector but are present in another.
- False Negative (FN): Records that the system failed to find.

The evaluation metrics considered are

- Accuracy provides the ratio of the number of right outcomes to the total number of inputs.
- Precision determines the percentage of correctly detected positives.
- Recall provides the percentage of genuine positives that were accurately recognized
- To calculate the score, the F1 Score considers both precision and recall. A harmonic-mean of recall and accuracy is used to get the F1 score. The best value is 1 and the worst value is 0, according to the F1-score.

Results and Conclusion

A total of 200 Facebook profiles were selected and required information was extracted from them. To these 30 artificially generated clones were added and fed to the detection modules to check how effectively they detect the clones. The modules performed as expected and detected clones accurately. The performance evaluation of clone identification using the Similarity measures module and utilizing the C4.5 module respectively, is shown in Tables 1 and 2.

Table 1: Performance evaluation of clone detection using similarity measures

Total no. of records checked	**230**
No. of normal records detected by system as normal (TN)	190
No. of normal records detected by system as clone (FN)	7
No. of clone records detected by system as normal (FP)	6
No. of clone records detected by system as clone (TP)	27

Table 2: Performance evaluation of clone detection using C4.5

Total no. of records checked	**230**
No. of normal records detected by system as normal (TN)	183
No. of normal records detected by system as clone (FN)	12
No. of clone records detected by system as normal (FP)	11
No. of clone records detected by system as clone (TP)	24

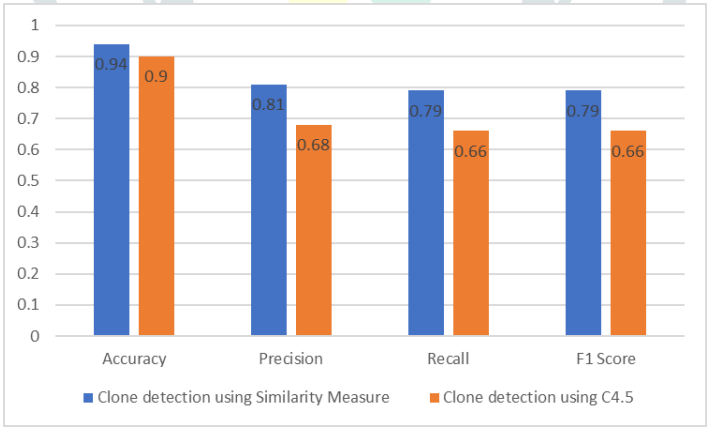

Fig 2: Performance Evaluation Result

Fig. 2. shows the performance evaluation result of clone detection using Similarity measures and C4.5 with respect to accuracy, precision, recall and F1-score. According to results, the Similarity measures module was able to accurately identify 27 out of 30 clone profiles, whereas the C4.5 module was only able to identify 24 out of 30 clone profiles. Therefore, it can be said that similarity measures are more effective for clone identification than the C4.5 classification technique.

REFERENCES

[1] Sowmya P and Madhumita Chatterjee, "Detection of Fake and Cloned Profiles in Online Social Networks", Proceedings 2019: Conference on Technologies for Future Cities (CTFC)

[2] Gordhan Jethava1 and Udai Pratap Rao, "A novel defense mechanism to protect users from profile cloning attack on Online Social Networks (OSNs)", Peer-to-Peer Network Applications, Springer Nature, 2022

[3] S Revathi and Dr M Suriakala, “Profile Similarity Communication Matching Approaches for Detection of Duplicate Profiles in Online Social Network”, 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), IEEE, Bengaluru, India

[4] Georgios Kontaxis, Iasonas Polakis, Sotiris Ioannidis and Evangelos P Markatos, “Detecting Social Network Profile Cloning”, 2013

[5] Piotr Brodka, Mateusz Sobas and Henric Johnson, “Profile Cloning Detection in Social Networks”, 2014 European Network Intelligence Conference

[6] Kiruthiga S, Kola Sujatha P and Kannan A, “Detecting Cloning Attack in Social Networks Using Classification and Clustering Techniques” 2014 International Conference on Recent Trends in Information Technology