International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 08 | Aug 2021 www.irjet.net p-ISSN: 2395-0072

PHISHING BASED WEB CONTENT MINING WITH LEAF CLASSIFICATION UNIT FROM DOM TREE

Ms. Paavai J, Ms. Saranya. D

*[*1]Research Scholar Department of Computer Science, D.K.M College for Women (Autonomous), Vellore. Tamilnadu, India.*
*[*2]Assistant Professor, Department of Computer Science, D.K.M College for Women (Autonomous), Vellore. Tamilnadu, India.*

Abstract: Online social networks (OSNs) slowly incorporate financial capacities by empowering the use of genuineandvirtualcurrency.Theyserveasnewplatformstoh ostavarietyofbusinessactivities,suchasonline headway events, where customers can get virtual money as prizes by participating in such events. Both OSNs and business accessories are significantly concerned when aggressors instrument a game plan of records to assemble virtual cash from these events, which make these events lacking and result in significant financial disaster. It is the destiny of unprecedented centrality to proactively separating these harmful records already the onlinepromotionactivitiesandsubsequentlydecreasestheir prioritytoberewarded.Inthispaper,wepropose a novel system, to be particular ProGuard, to accomplish this objective by proficiently planning features that depict accounts from three perspectives including their general practices, their invigorating cases, and the usage of their cash. We have performed expansive tests in light of data assembled from the Tencent QQ, an overall driving OSN with worked in financial administration exercises. Test results have exhibited that our framework can achieve a high recognition rate of 96.67% at a low false positive rate of 0.3%.

INTRODUCTION

BACKGROUND

An Online Social Network (OSN) is an accumulation of remote portable Accounts shaping a transitory system with no settled foundation or brought together specialist. In an OSN, every remote versatile Account works as an end-framework, as well as a switch to forward bundles. The Accounts are allowed to move about and sort out themselves into a system. OSN does not require any settled foundation, for example, base stations; in this manner, it is an alluring systems administration alternative for associating cell phones rapidly and immediately. For example, specialists on call at a catastrophe site or warriors in a war zone must give their own correspondences. An OSN is a conceivable answer for this need to rapidly build up correspondences in a versatile, transient and foundation less condition. This is one of numerous applications where OSN can be utilized. Versatile specially appointed systems are the eventual fate of remote systems. Records in these systems will create both client and application activity and perform different system capacities.

In the most recent decade, wired and remote PC organize upheaval has changed the processing situation. The potential outcomes and openings because of this insurgency are boundless; sadly, so too are the dangers and odds of assaults because of Fake Account by Fake Accounts. Counterfeit Account is characterized as an assault or a ponder un approved endeavor to get to data, control data, or render a framework untrustworthy or unusable. As indicated by, Threat can be characterized as "the potential probability of a consider unapproved endeavor to an

a) Access data,

b) Manipulate data

c) Render a framework temperamental or unusable.

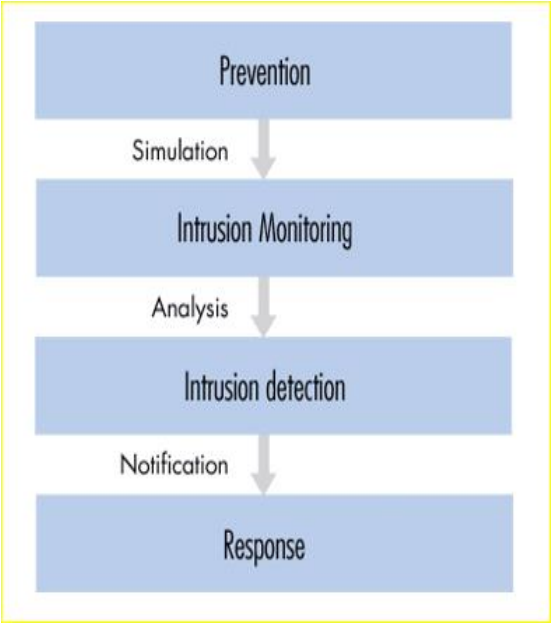

Fig 1: Fake Account Detection System Activities

Fake Account Detection Approaches

A Fake Account recognition is characterized as the strategy to recognize "any arrangement of activities that endeavor to trade off the honesty, secrecy, or accessibility of an asset". It is relating to procedures that endeavor to recognize Fake Account into a PC or a system by perception of activities, security logs, or review information. Review information will be data, which any Fake Account identification plan can deal with to decide whether any Fake Account has happened. The review information might be acquired on the host, in the application or the framework log record through host based Fake Account discovery framework or from the system through system based Fake Account identification framework. Counterfeit Account Detection Approach (IDA) fills in as a caution instrument for a PC framework or system. It recognizes the security bargains that happen to a PC system or framework and after that issues a caution message to an element, for example, a site security officer so the element can take a few activities against the Fake Account. An IDA contains a review information accumulation specialist, which monitor the exercises inside the framework, an indicator which breaks down the review information and issues a yield answer to the site security officer.

A security is an essential worry for OSN. To address these security related issues, there is a requirement for a proficient and successful Fake Account identification and reaction system that could recognize and react to the security related assaults at an Account level for OSN. This exploration work endeavors to address this need. So the issue articulation for the thesis can be expressed as takes after: To plan a Fake Account discovery and reaction security show, which would recognize and control the security related assaults at the Account level for Online Social Network in remote condition.

- The goal of this examination is to grow such a proficient and compelling Fake Account location and reaction system with the accompanying highlights:
- Identify the assault touchy system parameters and their edge esteems to develop the discovery and reaction models.
- Detect risk at the Account level viably in the OSN condition.
- Identify the gatecrasher that caused the risk.
- Respond to the gatecrasher and assaults, accordingly controlling the assault and securing the OSN.
- Design and develop the OSN Fake Account discovery and reaction structure with the end goal that it is convention autonomous.
- In this examination work to outline and built up a security demonstrate called the Fake Account Detection and Response for Online Social Network with the accompanying primary commitments:
- Identification of critical assault touchy parameters through machine learning based choice trees idea; recognizable proof of their edge esteems utilizing six sigma idea to separate their ordinary, questionable and powerless states for their application in OSN Fake Account identification and reaction.
- Formulation of the measure called Threat Index for powerful Fake Account Detection on OSN. Risk Index is processed utilizing fluffy rationale.
- Intruder distinguishing proof and reaction system display for assault control and insurance of OSN portable Accounts that are under risk by recognizing the interloper and subjecting fitting reaction plan.
- Protocol autonomous foundation for shielding the OSN from dynamic assaults by estimating basic parameters in the basic OSN framework. The proposed show ceaselessly screens the online system information and proficiently distinguishes the attacks, irrespective of the protocol used in OSN.

LITERATURE SURVEY

Classification and Review of OSN Security Schemes

In this area, we group the OSN security work into four general classifications in light of the sort of assault: validation, foreswearing of administration, narrow minded Account, and steering. In specially appointed systems, a versatile Account or host may rely upon different Account(s) to course or forward a parcel to its goal. The security of these Accounts could be imperiled by an outer assailant or because of the narrow minded nature of different Accounts. This would make an extreme danger of Denial of Service (DoS) and steering assaults where Fake Accounts join and deny the administrations to authentic Accounts. Not at all like Accounts in a wired system, the Accounts of OSN may have less handling power and in addition battery life and thus would endeavor to save assets. In this situation, the standard confirmation and encryption techniques would not make a difference to an OSN a similar way they would in a wired system. Notwithstanding, both verification and encryption

are much more imperative in an OSN. Steiner et al have built up a Group key Diffie-Hellman (GDH) show that gives an adaptable answer for assemble key administration. Yi et al have built up the MOCA (Mobile Certification Authority) convention that oversees heterogeneous portable Accounts as a feature of an OSN. MOCA utilizes Public Key Infrastructure (PKI) innovation.

The effect of verification assaults is very across the board and it incorporates unapproved get to, foreswearing of administration, disguising, data spillage, and space capturing. Capkun et al have built up a few arrangements utilizing an idea that they present, called Maximum Degree Algorithm (MDA), for anticipating disavowal of administration because of poor key administration. Avoine et al have built up a cryptography-based reasonable key trade display called Guardian Angel. This model uses a probabilistic approach without including a confided in outsider in key trade.

RELATED WORK:

A machine learning systems, including OSN Networks, Fuzzy Logic, MAD Algorithms, and so on have been utilized on KDD CUP 1999 information for Fake Account Detection with OSN arranges as principle apparatus in this sort of issue. Distinctive OSN arrange calculations have been utilized, including Gray OSN Networks, RBF Recirculation OSN Networks, PCA and MLP, with MLP for the most part indicating preferable outcomes over others. These works are primarily concentrating on abuse location. Keeping in mind the end goal to consolidate abuse and irregularity recognition, numerous scientists have as of late endeavored cross breed techniques, by joining OSN systems with other machine learning components, for example, fluffy rationale has a tendency to be better instrument of bunching, as it is speedier and more appropriate for constant frameworks.

A MAD is ordered as conduct based framework, when it utilizes data about the typical conduct of the framework it screens. Conduct on location depicts the reaction of the MAD after the discovery of assaults. It very well may be separated into dynamic or aloof in view of the assault reaction. These two kinds of Fake Account identification frameworks vary altogether from each other, however supplement each other well. The design of host-construct is totally reliant with respect to specialist based, which implies that a product operator lives on every one of the hosts, and will be represented by the principle framework. Furthermore, more proficient host-based Fake Account recognition frameworks are fit for checking and gathering framework review trails progressively and in addition on a booked premise, along these lines appropriating both CPU usage and system overhead and accommodating an adaptable methods for security organization.

Fuzzy Based Approach to Detect Black Hole Attack

Fluffy rationale is a numerical worldview to manage vulnerability about the information that complicated in the human elucidation. It communicates any announcement in phonetic way which makes fluffy manage based frameworks dazzling for application. have proposed Fake Account discovery framework which is fluffy rationale based framework to recognize dark gap assault on AODV convention in portable specially appointed system and it tends to different identification strategies in view of just a single factor for location reason and there are additionally some recognition frameworks which utilizes brought together way to deal with distinguish the Fake Account. Be that as it may, the proposed framework distinguishes the Fake Account through two factors, for example, goal grouping number, and forward parcel proportion which will discover all the delegate Accounts to achieve the goal and send bundles to middle of the road Accounts toward the start of transmission. On the off chance that the delegate Account neglects to send parcels then it sends test message to next Account since OSN is multi bounce in nature. At that point the framework fuzzifies the conveyance proportion on each neighbor bounce. It reviews the reaction time or affirmation time for every mediator Account and in light of which the Account will be identified as assaulted Account or something else. The test messages won't be sent by inert Accounts. The disadvantage of this technique is that it identifies just dark gap assault. Vitality Based Trust Solution For Detecting Selfish Accounts In OSN Using Fuzzy Logic

This framework will decide if the doubted Account is entirely a Fake Account or not. The proposed framework has four modules to recognize the Fake Account which are director, aggregator, trust adding machine and disseminator. In manager module, neighbors will be checked with the assistance of PACK (detached affirmation) framework that break down whether the Accounts extremely forward the bundles or not through acutely tuning in to their correspondence. On the off chance that there is any deviation from ordinary conduct it summons a total module. This module computes the quantity of parcels dropped by Accounts. And after that the fluffy based trust esteem will be computed for Accounts in fluffy trust module. It has three parts to figure the trust estimation of each Account - Direct trust esteem ascertained by coordinate trust operator (DTA), aberrant trust esteem computed by roundabout trust specialist (IDTA), aggregator which utilizes DTA and IDAT to figure the aggregate trust esteem. Since fluffy rationale gives exact outcome, the proposed framework utilizes fluffy rationale to figure the trust estimation of the objective Account. It likewise chooses the trust esteem in view of one single participation work. The disadvantage of this framework is the multifaceted nature of calculation and it

finds just the childish Accounts in the system. This technique isn't material to anticipate the sort of assault.

PROPOSED METHODOLOGY

A. Data Collection Methodology

In our work, we needed real world Facebook datasets, which are not available publicly. Some social graph datasets are available, which include profile-based feature data as well, but such datasets are in anonymized form and as such could not be used. Therefore, there was a need for us to collect data from Facebook API, which however is very restrictive with respect to the data that can be collected due to privacy issues. Authors working on Facebook like [10] often cite these difficulties. In order to characterize real and fake users on Facebook, we needed to collect data related to them, therefore, we adopted various strategies for such a data collection, given next are ways to collect ground truth data for real users as well as fake users. Table 1 summarizes the data collected. We will discuss the criteria for labeling the datasets in the following subsections. Facebook Real User Ground Truth: In order to gather real user data, we collected data comprising of user-feeds for users starting with authors(our own)social neighborhood. This includes the users at a single hop distance (friends) from our own node, that is, the direct friends appearing in our friend

TABLE I: Data Collection and Annotation

Description	Value
Total users for which data is collected	4,708
Number of Real Users	549
Number of Fake Users	230
Number of Users Assumed As Real	2,672
Number of Users Assumed As Fake	1,257

list, as well as friends of our friends (upto a distance of two) following a Breadth-First Search (BFS) traversal. For annotation, we consider all our immediate friends (549 in total) as real users and friends of our immediate friends (2672 in total) as assumed real users

.Facebook Fake User Ground Truth: In order to capture fake user behavior, we collected data for two types of users namely spammers and black market users. We manually identified 230 user accounts on Facebook who were involved in spamming activities on public Facebook pages, we annotate them as fake users. They would typically spread a lot of promotional messages and other spam content. In addition, there are many black market services available3. We registered and availed their services through which we found another set of 1257 users who we annotate as assumed fake users because we were not sure about their being fake.

B. Feature Identification

After collecting the various data attributes, next step was identifying and defining a set of features derived from these data attributes that would help in distinguishing real users and fake users as far as possible.

C. Learning Classifiers

As a final step in our methodology for detection of fake accounts on Facebook, we applied supervised machine learning classification algorithms. Supervised learners take annotated datasets as input and construct predictive models, which are used for tasks that involve prediction of one value using other values in the dataset. The two classes in our case are real users and fake users. The premise of using learning classifiers (as is the approach followed by many other authors mentioned in related work)is that the values of features recorded over long term are likely to be different for real user accounts and fake accounts, which are involved in various anomalous activities. A total of 12 supervised machine learning classification algorithms were used (from Weka), namely, k-Nearest Neighbor, Nave Bayes, Decision Tree classifiers (J48, C5.0, Reduced Error Pruning Trees Classification (REPT), Random Tree, Random Forest),

EVALUATION

In this section, we describe the evaluation methodology that we used to determine the performance of detection of fake accounts on Facebook .All the classifiers mentioned above were

applied to a mixed dataset, composed of a priori known Real accounts belonging to the first and second level of users in our social neighborhood (548 accounts), as well as the Fake accounts belonging to the set of spam accounts (229 accounts). The dataset also included the accounts belonging to users who are friends of friends in our social neighborhood, assumed

as Real (523 accounts), as also the black-market accounts, assumed as Fake (324 accounts).We evaluated the ability of various machine learning classification models in making predictions for unknown user accounts. The classifier implementations in Weka were used for this purpose. Two types of cross validation were performed, namely, the holdout method, and 10-fold cross validation. Given below are the various evaluation strategies that were adopted along with their results.

A. Performance Evaluation of Learning Classifiers

The aim of this evaluation was to understand which among the 12 learning classifiers perform the best. As a baseline, in each experiment for each learning classifier, 10-fold cross validation was done on a dataset of 777 user

accounts comprising of 548 real accounts and 229 fake accounts belonging to the set of spam accounts. The entire feature set comprising of 17 features were employed. Results are depicted in Fig. 1 in which it is evident that classifier accuracy of close to 80% is achievable using conventional learning classifiers

Machine Learning Classifier

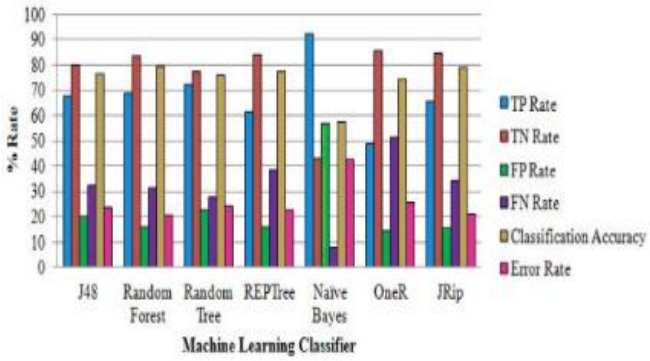

Fig. 1: Performance Results of Prominent Learning Classifiers

Only those learning classifiers that performed relatively well with a detection rate of 75% or greater and an error rate of 25% or less are depicted in Fig 1. Key observations are (1) The decision tree and decision rules classifiers were the ones that performed well. (2) All the 7 classifiers with the exception of Nave Bayes gave results with detection accuracy in the range 75-80% and an error rate of 20-25%. (3) Nave Bayes resulted into the lowest accuracy rate of 58% but could detect 92% of the actual fake profiles, which is the highest of them all. (4) Apart from this, with the exception of One R, all the classifiers were able to detect 62-73% of the actual fake profiles.

B. Performance Evaluation of Feature

In this evaluation, we divide entire feature set into various categories on the basis of Facebook activity type. We aim to determine the ability of each category (set) of attributes, to distinguish fake users from real users. We performed a 10-fold cross validation on the same dataset as used earlier. Although the evaluation was done for all the learning classifiers, only two are depicted namely J48 in Fig 2 and Random Forest in 3, respectively

J48 Classifier

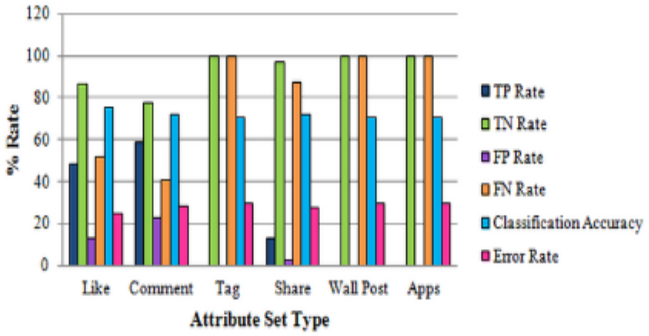

Fig. 2: Feature Level Performance Evaluation for J48 Classifiers

Random Forest Classification

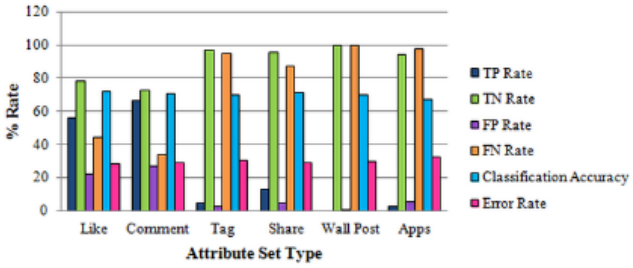

Fig. 3: Feature Level Performance Evaluation for Random Forest Classifiers

The results for all the well performing classifiers except Nave Bayes were found to be on similar, therefore, only the two classifiers are depicted. Key observations are (1) All the classifiers resulted into classification accuracy of 70%and above and error rate of 30% and below for all sets of attributes. (2) The comments-related and likes-related attributes were successful in detecting fake profiles, with a TP rate of 60% and above for comments-related attributes and a TP rate approaching 50% and above for likes-related attributes. (3) Attributes like share, tag, wall post and apps exhibit good performance in detecting real accounts, with TN rates of 95%.

(4) At the same time, these features could not detect fake profiles (TP rate 5% and below), a plausible reason could be that there is not enough of such activities done by fake accounts (spammers in particular) in our dataset. Results for Naive Bayes were different, so they are specially depicted in Fig 4.Key observations are: (1) Although the likes and comments related attributes result in a detection accuracy of 70% and

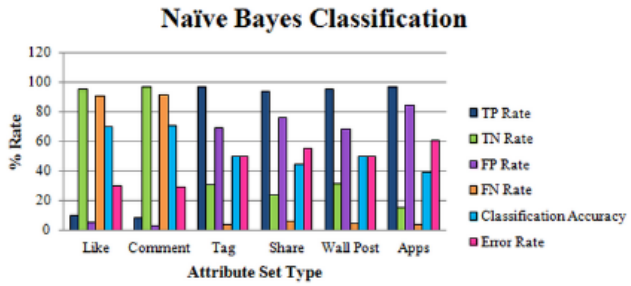

Fig. 4: Feature Level Performance Evaluation for Naïve Bayes Classification

Error rate of 30% and below, similar to the results for other classifiers, but for other sets of attributes, the accuracy is 50%and below and error rate is high, 50% and above. (2) Also, the TP rate for likes and comments features are very low (below 10%), while those for tags, share, wall post and apps related attributes are very high 95% and above. On the contrary, the TN rate for likes and comments related attributes is 95% and above, and for the remaining attributes, it is 30% and below. (3) In addition, the likes and comments related attributes result in a FP rate of 5% and below and FN rate of above 90%. On the other hand, for the remaining attributes, the FP rate and FN rate are 70-85% and 0-6%, respectively. A possible explanation of these findings is that Nave Bayes is highly dependent on past (or already observed) phenomena and fake accounts consciously make efforts to align their post activities closer to that of real users. Thus, the real-like behavior makes it difficult for Nave Bayes to correctly detect fake accounts

C. Prediction of Unknown User Accounts

This evaluation was performed on the entire dataset. The training set for this experiment comprised of 777 accounts with 548 real accounts and 229 fake accounts. The aim of this evaluation as depicted in Fig. 5 was to predict labels (Real/Fake) for the accounts in the test set, which were labeled on the basis of assumptions and analyze the results.

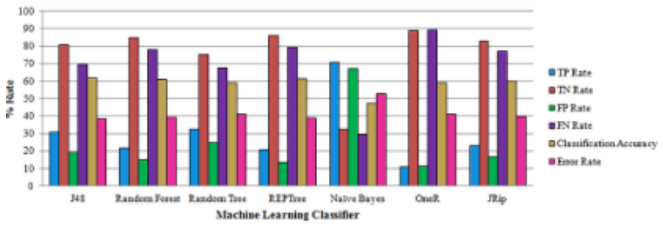

Fig. 5: Prediction Performance of Learning Classifiers

The key observation here is that all the classifiers except Nave Bayes gave a detection accuracy of 60% and error rate of 40%. These results are not very promising, however, the work can be treated as a baseline to make further improvements in future.

CONCULSION

Fake accounts have been constantly evolving over the years in order to evade their detection. Thus, it is important to develop techniques for detecting fake accounts, keeping into account their near-real behavior. In response to this requirement, we have made our first attempts to detect fake accounts on Facebook based on the user profile activities and interaction with other users on Facebook. These activities were characterized through an exhaustive feature set covering the like, comment, share, tag habits and apps usage of Facebook users.This research designed and developed the Fake Account detection and response model for Online Social Networks (OSMAD). This model consists of the Fake Account detection framework and the Fake Account response framework. These two frameworks complement each other to make a complete Fake Account detection based security model for OSNs. The functionality and effectiveness of this model was validated by applying this model for simulated Denial of Service attacks (DoS) in OSN. The detection framework of the OSMAD uses CART based data mining methodology to identify the parameters that are significant for a particular attack. It then uses the six sigma methodology to set the thresholds for the significant parameters identified. The detection framework then quantifies an attack using a metric called Threat Index (TI) by applying fuzzy logic on the measured values of the significant parameters. The response framework of the OSMAD is invoked when the detection framework identifies an attack. The response framework has an intruder identification component and an Fake Account response component. The intruder identification component uses the significant parameters and the thresholds in a reputation management mechanism to identify the intruder and flag its status. The response action plan is then executed by the response framework based on the intruder status indicated by the reputation management mechanism.

FUTURE WORK

As part of our future work, we would like to work on a number of aspects of fake accounts detection on Facebook.

(1)Improving Labeling Criteria: We would like to design and test alternative labeling techniques for the user accounts.

(2) Feature Set Improvement : We plan to refine our existing feature set to incorporate inter-user behavioral pattern with an aim to further improve the accuracy of fake accounts detection.

(3)Improving Detection Accuracies: We also intend to apply more classification algorithms on larger datasets to further improve our detection accuracies.

(4)Graph-Based Techniques: We may also formulate a hybrid approach for detecting fake profiles by combining our approach with graph-based techniques.

(5) Online Application and Mobile Application: We would like to provide a web service to Facebook users as an online application or a mobile application to facilitate determination of a given profile as fake or real.

(6)Real Time Detection : We may also develop a browser extension that can work as a detector of fake accounts in real time. We acknowledge that for real time detection, the collection of labeled data generated in this work may become obsolete in the longer term. Thus, we may work on periodic learning techniques by getting periodic label feedback from people.

Since not many research efforts have been devoted to OSN IDA, especially the intruder identification, Fake Account response and control, this dissertation provides the leading effort in constructing a viable OSN Fake Account detection, Fake Account response and control model. As a very new, hot and promising research area there are several interesting and important future directions explained as follows.

- Focusing on DSDV and AODV as the routing protocol, and DoS attacks as the threat model, we have designed and developed OSMAD model to the full. Further work can be performed to extend this model to other passive attacks like unauthorized access, probing, selfishness and non-repudiation attacks in Online Social Networks and active routing attacks.

- Further research could also be devoted to apply the proposed model to secure the integrated wired and wireless networks like cellular networks/sensor networks.

- Having demonstrated the viability of Fake Account detection and response approach in providing security for Online Social Networks in a simulation environment, it would be valuable to evaluate the model in a wireless OSN test bed in order to bridge the gap between simulation and actual OSN deployment.

REFERENCES:

[1] Y. Li and J. Wei., "Guidelines on selecting Fake Account detection methods in OSN", In Proceedings of the Information Systems Educators Conference, 2004.

[2] Bo Sun and Lawrence Osborne: Fake Account detection techniques in mobile ad hoc and wireless sensor network. In: IEEE Wireless Communications,1536-1284, 2007.

[3] R. Heady, G. Luger, A. Maccabe, and M. Servilla, "The architecture of a network level Fake Account detection system" Technical National Conference on Emerging Trends and Applications in Computer report, Computer Science Department, University of New Mexico, August 1990.

[4] B. Shanmugam and N. B. Idris, "Anomaly Fake Account Detection based on Fuzzy Logic and Data Mining", In Proceedings of the Postgraduate Annual Research Seminar, Malaysia 2006.

[5] M. Wahengbam and N. Marchang, "Fake Account detection in OSN using fuzzy logic", 3rd IEEE Science (NCETACS), ISBN: 978-1-4577-0749-0, pp. 189 – 192, Shillong, 30-31 March 2012.

[6] S. Sujatha, P. Vivekanandan, A. Kannan, "Fuzzy logic controller based Fake Account handling system for Online Social Networks ", Asian Journal of Information Technology, ISSN: 1682- 3915, pp.175-182, 2008.

[7] S. Ahmed & S.M. Nirkhi, "A Fuzzy approach for forensic analysis of DDoS attack in OSN" International Conference on Computer Science and Information Technology, ISBN: 978-93-82208-70-9, Hyderabad, 10th March 2013.

[8] VydekiDharmar and R.S. Bhuvaneswaran, "A combinatorial approach for design of fuzzy based Fake Account detection system", proc. of international conference on computer applications (ICCA) 2012.

[9] M. B. Mukesh Krishnan, P. Sheik Abdul Khader, "Fuzzy Based Integrated Security Model for Mobile Ad Hoc Network", Global Trends in Computing and Communication Systems Communications in Computer and Information Science Volume 269, 2012, pp 467-

[10] Chaudhary, A., Kumar, A., & Tiwari, V. N. (2014, February), " A reliable solution against Packet dropping attack due to Fake Accounts using fuzzy Logic in OSNs", In Optimization, Reliability, and Information Technology (ICROIT), 2014 International Conference on (pp. 178-181), IEEE.

[11] Chaudhary, A., Tiwari, V. N., & Kumar, A. (2014, February), "Design an anomaly based fuzzy Fake Account detection system for packet dropping attack in mobile ad hoc network", In Advance Computing Conference (IACC), 2014 IEEE International (pp. 256-261), IEEE.

[12] Chaudhary, A., Tiwari, V. N., & Kumar, A. (2014, February), "Design an Anomaly Based Novel Approach for Detection of Sleep Deprivation Attack in Online Social Networks Using Soft Computing", Proceedings of 3rd International Conference on Recent Trends in Engineering & Technology (ICRTET'2014), Elsevier.