IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON INTELLIGENT SENSING ON MOBILE AND SOCIAL MEDIA ANALYTICS

Received January 1, 2017, accepted January 18, 2017, date of publication March 15, 2017, date of current version March 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2666823

ISMA: Intelligent Sensing Model for Anomalies Detection in Cross Platform OSNs With a Case Study on IoT

VISHAL SHARMA[1], ILSUN YOU[1], (Senior Member, IEEE), AND RAVINDER KUMAR[2]
[1]Department of Information Security Engineering, Soonchunhyang University, Asan-si 31538, South Korea
[2]Computer Science and Engineering Department, Thapar University, Patiala 147004, India

Corresponding author: I. You (ilsunu@gmail.com)

This work was supported in part by the Soonchunhyang University Research Fund and in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2016R1D1A1B03935619.
This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author.

ABSTRACT In the recent years, the user activities over online social networks (OSNs) have increased tremendously. A large number of users share information across the different social networking platforms. The information across the OSNs is easy to access, and thus, can be easily used by the fraudulent users for misleading the entire community. Such fraudulent users are termed anomalies. In this paper, a problem of cross-platform anomalies is considered, which possesses different behaviors by an individual with different users across the multiple OSNs. The variation in the behavior and activity makes it difficult to identify such anomalies. A solution to this problem is proposed on the basis of cognitive tokens, which provide an intelligent sensing model for anomalies detection (ISMA) by deliberately inducing faulty data to attract the anomalous users. A common login system for different OSNs is also suggested as a part of collaborative anomaly identification across different OSNs. A fair play point approach is used for the determination of anomalies. Both simulations and email-based real data sets are used to measure the performance of the proposed approach. Furthermore, as an example of implementation, a case study is presented for anomaly detection in Internet of Things. The proposed approach is able to provide the highest accuracy at the rate of 99.2%; this is 25.1% higher as compared with the SVM-RBF and sigmoid approach, and 22% higher than that of the k-nearest neighbor approach. Furthermore, the proposed ISMA also caused less error in detecting the anomalies, which were within the range of 0.1% to 2.8%. The error in identification is reduced up to 96.6% in comparison with the SVM and k-nearest neighbor approaches. The gains in comparative results validate the efficiency of the ISMA in identification and classification of anomalies in cross-platform OSNs.

INDEX TERMS OSNs, anomalies, intelligent sensing, IoT, cross-platform.

I. INTRODUCTION

With an increase in the demand of interactions between the users, Online Social Networks (OSNs) are playing a crucial role in connecting the users. OSNs include social networking sites which aim at bridging the gap between the users across the globe. The ease of connectivity and access to all the data allows gathering a huge amount of information about various objects and users [1]. OSNs have been a rapid source of information exchange where a huge amount of data transactions takes place within few seconds. OSNs can operate independently or can have multiple dependencies for data gathering using web crawlers [2]. Despite the advantages of OSNs, there is always a risk of information loss or access to misleading users. The faulty users can utilize the information access through OSNs to cause a devastating effect to online traffic as well as its infrastructure. Anomalies are one of the crucial issues for cyber security as consistent monitoring and analyses of accounts are required to identify them. Anomalies can be of different types with different properties. An anomaly may be a vertical threat caused by a user to the OSNs' servers or can be horizontal threat caused to fellow users [3], [4]. Anomalies arise with the sudden variations in the activities of users or sources which keep on changing with the passage of time. An illustration of the anomalous

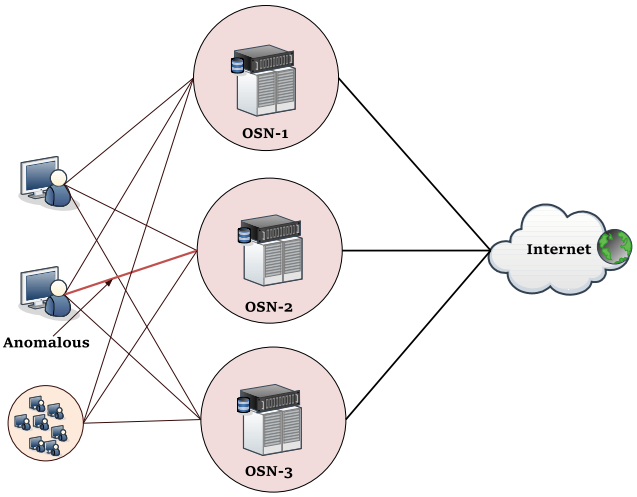

FIGURE 1. An illustration of OSNs with an anomalous user.

user across multiple OSNs is shown in Fig. 1. The change in activities can be the variations in the pattern of accessing the information or causing impact to already existing fraudulent information [5], [6]. The detection and prevention of the anomalies are one of the major challenges for the current OSNs. One of the typical anomalies which are difficult to identify and classify is the anomalies across the cross-platform OSNs [7]. With different types of OSNs, a single user may have accounts with almost all the OSN platforms, and its activity may vary for different OSNs. For one OSN, a user may be a silent information seeker whereas for other it may access the viable information for manipulation. Such anomalies possess different behavior with different types of sources across the different platforms and are classified as ''Horizontal Anomalies'' [1]. Horizontal anomalies include variations in the activity amongst the fellow users; whereas as the vertical anomalies include a threat to the underlying infrastructure by its users. A cross-platform anomaly mainly involves different social networks acting as different sources for a common user, whereas a horizontal anomaly refers to the difference in behavior and activity of a user only on the basis of different sources with or without the involvement of multiple social network platforms.

The major characteristics of anomalies across the multiple OSNs involve the different level of interaction with different users, the difference in the type of information accessed, and the difference in the usage sessions. These anomalies have different patterns at a particular instance with different OSNs. Thus, classification of features needs to be done accurately for identification of such anomalies. The difference in patterns requires common features which can be applied to all the social networking sites, and thus, can be used to efficiently tackle the anomalies. With an appropriate solution, an anomaly in one OSN can be classified as a potential threat by the other OSNs. However, such solutions require common access IDs or user tracking systems, which can identify the same user across the multiple social networking platforms.

There are various existing solutions that aim at the identification and classification of anomalies by adopting supervised or unsupervised learning using the data collected over a period of time. Such approaches are often operated off-site at the data centers and potential anomalies are marked in a single social networking platform. Some of the popular solutions include outlier filter algorithms [8], [9] which utilize the density-based solutions to identify the anomalies; distance-based approach [10] which utilizes the concept of minimum distance to identify the anomalies, and regression approaches such as Support Vector Machine (SVM) [11]. Supervised or semi-supervised approaches can utilize the output trends to identify the anomalous data. However, the real-time usage of these approaches is difficult. Apart from that, the existing solutions cannot be used to detect anomalies across different social networking platforms at real time, which is considered as a problem in this paper.

Thus, aiming at efficiently identifying the anomalies across the cross-platform OSNs, an Intelligent Sensing Model for Anomalies (ISMA) is developed, which not only identifies the anomalies but also classifies them into communities. The proposed approach provides an efficient solution which operates in two parts. The first part finds the potential anomalies and the second part utilizes the potential anomaly data to finalize the actual anomalies in the multiple OSNs. The concept of cognitive tokens is used to identify potential anomalies and error-based outlier algorithm is used to classify them into communities. The evaluation of the proposed approach is presented using both synthetic as well as real-time email datasets. The proposed approach is also verified theoretically and statistically. Moreover, a state-of-the-art comparison is presented with the existing SVM [11] and k-nearest neighbor approach [10]. The proposed approach is able to provide an efficient detection and sensing defined in terms of the accuracy and error in finding anomalies.

Similar to OSNs, Internet of Things (IoT) is one of the important examples of cross-platform systems [12]. Device to device connectivity is one of the major advantages of IoT. With billions of devices connected to each other via gateways or as a direct connection, a scenario may arise with users having different patterns of interaction with the different devices [13]–[15]. Such scenarios raise the requirement of anomaly detection in IoT. Anomalies in information and communications can lead to several network attacks which may compromise the network information with the intruder [16], [17]. A case study is presented to prove the utility of the proposed approach for the identification of anomalies amongst IoT devices. The results presented in the case study suggest that the proposed approach can be used to find anomalies in IoT without much utilization of memory and with a high likelihood at every instance. Running time, which may increase with the increase in the number of anomalies, can be controlled by the parallel run of the proposed model. The key highlights of the work presented in the paper are:

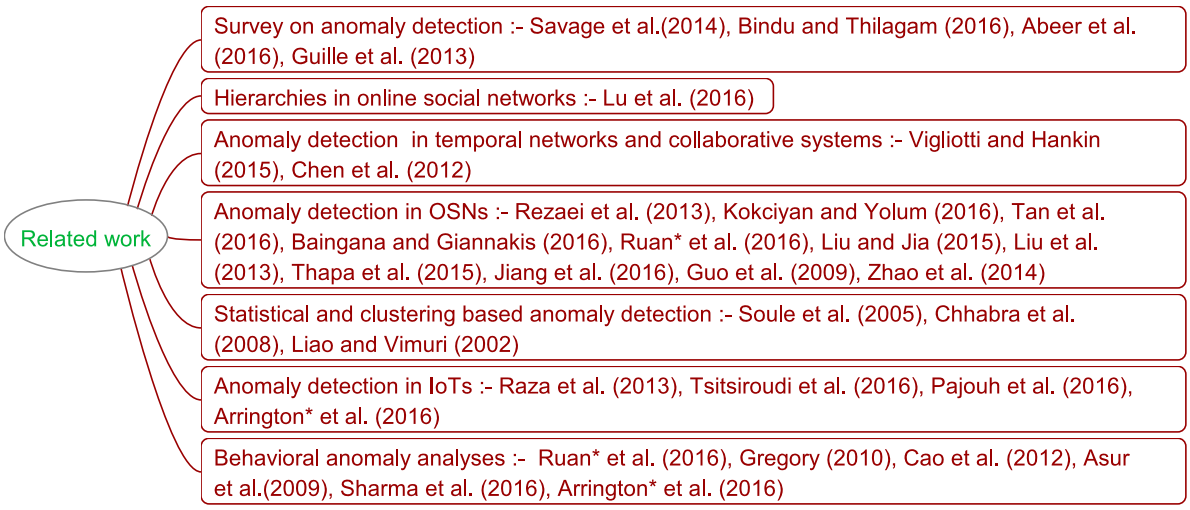

FIGURE 2. A broader view of existing solutions.*multiple categories.

- Formation of an intelligent sensing platform for detecting cross-platform anomalies.
- Implementation of proposed approach with existing outliers.
- Formation of a new outlier approach to enhance the efficiency of the proposed approach.
- Testing and validation by comparison with state-of-the-art approaches using real datasets.
- Analyses over IoT environment for anomaly detection.

The rest of the paper is structured as follows: Section II gives the insight into existing literature. Section III gives the motivation and problem statement. Section IV gives the preliminaries and definition of new terms. Section V presents the system model. Section VI gives the detailed proposed approach. Section VII validates the proposed approach theoretically. Section VIII presents the performance evaluation using synthetic datasets and statistical analyses. Section IX gives the state-of-the-comparison. Section X gives a case study and testing over IoT environment. Section XI highlights some open issues with conclusion presented in Section XII.

II. RELATED WORK

Anomalies have been a subject of concern in most of the research areas. A broader view of the related works considered in this paper is shown in Fig. 2. With more number of users sharing information over the OSNs, it has become easier for the intruders to gather this information and cause a devastating effect to users as well as their communities [18]. The rise in the number of anomalies can lead to theft of personal information that can compromise user privacy [19]. A lot of researches have been carried over the years which aim at the identification and elimination of anomalous users from the OSNs. Anomalies can be of different types, and the solutions are developed customarily to identify one type of anomaly on the basis of patterns for a particular feature. Using rumors as the basis, Tan *et al.* [20] developed an elastic collision model for the detection of anomalies. The authors concentrated on the rumor propagation across the OSN to identify the nodes which spread rumors. The approach can only be used for identification of fake sources. A solution to user tracking can be the collective effort made by all the communities. Considering the time varying networks, Baingana and Giannakis [21] developed an approach for the anomaly tracking using joint communities. The dominant community approach is utilized to identify the potential anomalies across the dynamic networks. Ruan *et al.* [22] aimed at the profiling of users accounts on the basis of social behaviors. The authors focused on determining the compromised accounts using social behavior characteristics. These approaches utilize a single network system for determining the anomalies while operating at different levels without considering the horizontal classification.

Anomalies detection approaches for collaborative systems can also be adapted for the identification of faulty users across the cross-platform OSNs. However, such approaches require remodeling to make themselves suitable for cross-platform online networks. Since these approaches are already modeled for collaborative systems, the similar methodology can be directly applied to OSNs [23]. Cyber attacks can be mitigated by the formation of reputation systems which aim at maintaining the trust level and prevents against real user attacks. Liu and Jia [24] and Liu *et al.* [25] developed a similar system which provides anomaly detection on the basis of trust modeling. However, the scope of their work is only limited to stand-alone online systems. The authors also presented a work on OSNs by exploiting the service architecture to form an optimal social trust path for selection of users. Proximity can be considered as the other measure for the formation of

security protocols for user matrix in OSNs [26]. However, such approach requires critical privacy mechanisms which make it difficult to sustain for networks with a large number of users and increasing proximity despite the non-anomalous activity.

Detection of the overlapping communities and anomaly detection on the basis of behavioral variations are the other solutions for deciding the fraudulent users in the OSNs. Community detection can help identifying the users and can be labeled on the basis of their activity [27]. Such approaches operate in reverse, i.e. first the communities are classified, and then the users are checked for anomalous activity similar to the density-based outliers. These approaches can be efficient only for off-site evaluation of data but not for the real-time system deployments. Some of the behavior approaches include trust evaluation in OSNs by Jiang *et al.* [28], Coupled behavior in a community application by Cao *et al.* [29], auction framework by Thapa *et al.* [30], evolutionary behavior graphs by Asur *et al.* [31]. Despite being influential, the existing approaches can be used in the environments exactly similar to the one for which these are proposed. None of the existing solutions aim at the resolution of cross-platform anomalies in the users operating with different intentions across multiple OSNs. Thus, an efficient solution is required, which not only provides the classification of communities, but, should be applicable to multiple platforms at the real time.

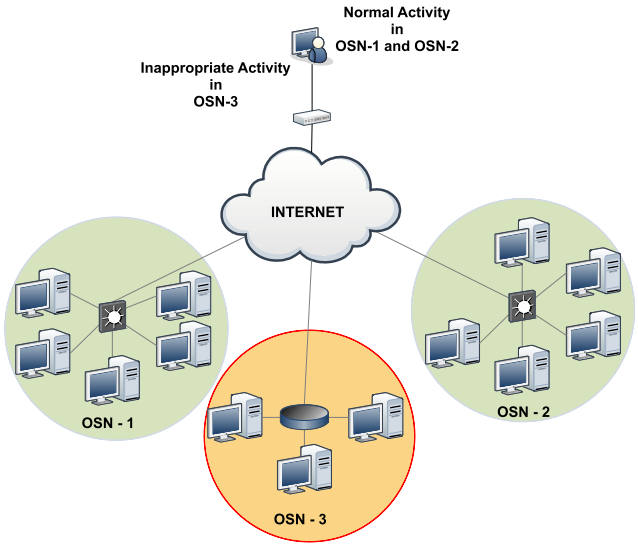

FIGURE 3. An illustration of difference in activity of a user across different social network platforms.

III. MOTIVATION AND PROBLEM STATEMENT

Cross platform OSNs are composed of multiple users who interact with different social networks such as facebook, twitter, g+, wiki, kaggle, researchgate, Linkedin, and youtube as well as make use of common or multiple accounts. A user with accounts on different social networks is termed as a cross platform user and the difference in the behavior of a user in different social networks is termed as a cross-platform anomaly, as shown in Fig. 3. The motivation of the research presented is derived from the viewpoint of identifying a user with the difference in activity towards different social networking platforms. The difference in the activity may or may not be abnormal, which is a problem to identify. The accurate classification of the user activity, mapping a user to the community, and then deciding whether the user activity leads to anomalous behavior or not is one of the major challenges in cross-platform social networks. A user with abnormal behavior in one type of social network is a potential risk for other social networks in which it pretends to be a normal user. Thus, the identification of such hazardous users with very less available information leads to the ideology of forming an intelligent sensing platform which can detect as well as visualize the cross-platform anomalies in OSNs.

Thus, the problem deals with the formation of an intelligent solution which can perform the tedious task of detecting anomalies across multiple OSNs. The major challenges involved in the formation of the proposed approach include,

- identification of a user who is an anomaly.
- identification of the community to which the anomalous user belongs.
- impact of anomalous user and community on the neighboring users and communities.
- sub-classification of the community on the basis of detected users.

Existing approaches involve classifying communities first, and then monitoring the activity of the user with respect to other users of the communities. This has been the popular approach and most of the existing density-based outliers use this policy to identify a potential anomaly. However, identification of a community may or may not classify a user which is an anomaly. This can be explained with the help of an example. Suppose there is a user X that belongs to two different communities, namely A and B. Now, using the common density-outlier, the user X is classified to community A as its properties match the properties required for being the A's user. Now, since X is having properties common to users in community A, there is always a high probability that the user X is having anomalous behavior to users in community B. If this is true, user X is a potential anomaly in between the two communities. Such behavior cannot be easily detected by the existing approaches. Thus, a novel and intelligent solution is required, which can classify the anomalous user first on the basis of their activities, and then monitor the entire community to which it belongs for detection of further anomalous users. Such detection and sensing is an example of cross-platform anomaly detection in OSNs.

Further, the upcoming demand of IoT also involves different devices which share information with multiple users having different characteristics. Such users are to be monitored and classified in the case of difference in activities with the different types of devices. Thus, the proposed solution is also presented as a case study in detecting anomalies amongst

different IoT devices depending on the activity classification of a user similar to the cross-platform OSNs.

IV. PRELIMINARIES AND TAXONOMY OF TERMS

The proposed approach aims at providing an intelligent solution for sensing anomalies in OSNs. This section provides insight into various terms and preliminary conditions for understanding the proposed approach. The network preliminaries are as follows:

- Online Social Networks (OSNs) and Anomalies: The online social network comprises users denoted by a set N which belong to a particular community denoted by a set C. Each user in a community can be connected to the n number of users across different social networks denoted by a set S. The entire network is subjected to set theory where S is the superset of different online social networks such that $S = \bigcup_{j=1}^{s} S_j$, where s is the number of social networks. Set C is the superset of communities such that $C = \bigcup_j \bigcup_i C_{i,j}$. Here, $1 \leq i \leq c$ and $1 \leq j \leq s$, where c is the number of communities in a single social network. The set $C_{1,1}$ is read as the set representing community i of a social network j. Each community comprises users from different social networks such that the overall set $N = \bigcup_p \bigcup_k N_{k,p}$, where $1 \leq k \leq n$ and $1 \leq p \leq c$. All the social networks are formulated into graphs $G_1, G_2, \ldots, G_s$, such that $G_s = (N_s, E_s, \pm W)$, where N_s is the number of users in a social network S_s with number of edges denoted by E_s, W is the weight for each edge and $\pm$ represents an in-degree or out-degree for the edge. The weights in a graph are decided on the basis of property such as friendship in the case of facebook, followers in the case of twitter, circles in the case of google plus, and collaborators in the case of researchgate, etc. A user of set N is said to be an anomaly if it possesses different behaviors and activities in the different social networks. The cross-platform anomaly is identified by the number of connections a user possesses in social networks.
- Dynamic and Static Social Graphs: Dynamic graphs are the graphs formed with the variation of a particular property acting as an edge between the nodes of a social network. A dynamic graph is labeled on the basis of a property which changes its value with the course of time or iteration. Usually, OSNs are expressed as the dynamic-labeled graphs to identify the behavior and activity of a user. Such graphs can be directly used to identify anomaly on the basis of variation in the properties by using approaches like time series, noise variations, etc. However, if a network remains unchanged for a longer duration of time, no variations are observed which makes it difficult to detect an anomaly using existing solutions. Static social graphs are the normal graphs which can be labeled or unlabeled with or without weights. Such graphs are formed by the vertices having edges which remain fixed during the entire period. Such graphs are easy to observe and operate for the identification of anomalies. However, the cross-platform social networks are observed as the dynamic labeled graphs with multiple edges existing between the nodes. Thus, the formation of a graph and selection of the approach play a key role in the detection of an anomaly in cross-platform OSNs. In the proposed approach, the initial graphs considered are the dynamic labeled social graphs which are converted into single-labeled social graphs with multiple instances. Multiple instances represent that for a same set of users one or more static graphs are observed on the basis of a number of properties considered for a link between nodes. Hence, if $G_s = \bigcup_i^m (N_{s,i}, E_{s,i}, \pm W_i)$ are the graphs for m number of properties, then the final static graph is given as $G_s = \bigcup_i^m (N_{s,i}, E_{s,i})$. However, the proposed approach utilizes the in-degree and out-degree concept since it has to capture anomalies on the basis of variation in the interaction with the different sources across multiple OSNs. Thus, in the proposed approach, the static graph formed is given as $G_s = \bigcup_i^m (N_{s,i}, E_{s,i}, \pm)$.

The definition of various terminologies used in the proposed are described as follows:

- Social strength (D_o): It refers to the total number of connections possessed by a user across all the OSN platforms such that $D_o = D_{o,1} + D_{o,2} + \ldots + D_{o,s}$. This is a measure of the total degree of a user. There is always a key challenge of detecting a user across multiple social networks which are mandatory to calculate the total degree. This issue is resolved by the novel ideology presented in the later part under the proposed approach.
- Distance of information retrieval (D_r): It refers to the direct connections of particular user across all social platforms such that $D_r = D_{r,1} + D_{r,2} + \ldots + D_{r,s}$. This helps to identify the reach of a particular user in a social network.
- Frequency of connections (F_c): It refers to the number of times a user interacts with other sources of a social network such that $F_c = F_{c,1} + F_{c,2} + \ldots + F_{c,s}$. This source can be another user, a server, a message hub, or a simple hyperlink. The frequency of connections decides the weight of an edge and also allows taking a decision of considering the edge or not. If F_c is too low, then the edge can be ignored and the dynamic labeled graph formed w.r.t. weight is neglected during the formation of a final static graph.
- Relation cost (R_c): It refers to the number of alternative paths available for a user to interact with the other users or sources across all the networks other than direct links such that $R_c = R_{c,1} + R_{c,2} + \ldots + R_{c,s}$.
- Satisfied trust level (S_{TL}): It is the score assigned to each user on the basis of its number of interactions with a false source, which is induced in the proposed approach in the form of intelligent tokens. S_{TL} accounts for the total number of hits/interactions made by a user with all the tokens floating across different OSNs.

- Threat level (T_L): It denotes the total number of different tokens encountered by a user during a particular course of time. This may range from one to many depending on the number of tokens floated in the OSNs.
- Activity time (A_t): It refers to the total active time of a user across all the social platforms such that $A_t = A_{t,1} + A_{t,2} + \ldots + A_{t,s}$. This helps in monitoring the time log of a user as well as it determines the number of dynamic labeled graphs formed considering the time interval between each formation.
- Fair play points (F_p): It is the user ranking system which helps to rank a user on the basis of its activity throughout the connectivity on the social network. This ranking system is similar to a third party maintaining a check on the user activity. The concept of fair play points is aligned with that used in many modern day sports, however, with a different formulation. In the proposed approach, the F_p is the social index value, which will be assigned to each user on the basis of its activity over different social networking platforms. This value can be displayed on the dashboards of users and may be displayed publicly to identify a potential threat to other users. Since F_p is derived over the number of personal activities of users across the online social networks, a short survey in the form of a questionnaire is conducted to check whether such public illustration and third party monitoring affect the user activity or not. A set of three questions were taken in the survey, namely,
  Q1. If social websites like Facebook, naver, twitter start giving a reputation score to each user; Will this affect your approach towards using these websites?
  Q2. Will displaying of social index restrain your activity on the internet?
  Q3. If you know that your online activities are being disclosed to a third party, will this affect your search activities?

TABLE 1. Survey for validation of selected parameters.

Answers	Q1	Q2	Q3
Yes	2	9	16
No	38	31	15
May/May not be	0	0	9

This survey was taken by 40 regular internet users and active members of online social media having accounts on a majority of the online social networking sites. The majority of them answered "No" to all the questions; however, some were worried about their privacy while few were unsure. The details of the survey are presented in Table 1. From this survey, it can be analyzed that the proposed approach which utilizes all the metrics explained above can be successful by gathering data at the will of a user.

V. SYSTEM MODELING

The proposed approach aims at the formation of an intelligent sensing system for the detection of cross-platform anomalies. The model utilizes the exponential growth formulation to determine the user activity across different social networks [32]. The entire system is observed w.r.t. max-min conditions, which on satisfaction gives accurate results for anomaly detection. For non-anomalous behavior,

$$\begin{aligned} D_o &= \max, \\ D_r &= \min, \\ F_c &= \max, \\ R_c &= \min, \\ S_{TL} &= \min, \\ T_L &= \min, \\ F_p &= \max. \end{aligned} \tag{1}$$

The exponential growth model helps in understanding the alterations in the value of F_p, which is a crucial measure for a user being an anomaly or not. According to exponential growth,

$$F_{p,A_t} = F_{p,A_{t-1}} (1 + I)^{A_t}, \tag{2}$$

where $F_{p,A_{t-1}}$ is the previous fair play point for a user, I is the rate given as:

$$I = \frac{D_o}{R_c T_L}. \tag{3}$$

The entire system depends on the Error in fair-play (E_g) value for determining the anomaly in the system of multiple social networks. E_g forms the basis of the novel error-based outlier, which is used to determine the final set of anomalies in the multiple social networks along with the communities to which an anomalous user belong, and the communities which may be affected due to the detected anomalous user. In the proposed model, error in fair-play is given as:

$$E_g = \frac{f(B, A_t)}{F_p}, \tag{4}$$

where $f(B, A_t)$ is the nonlinear function derived from Michaelis-Menten Enzyme (MME) model [33]. Using MME, the function is defined as:

$$f(B, A_t) = \frac{B_1 A_t}{B_2 + A_t}. \tag{5}$$

For simplification, $B_2 = \frac{B_1}{2}$, and

$$B_1 = D_r e^{-D_r S_{TL}}, \tag{6}$$

such that Eq.(5) is given as:

$$E_g = \frac{2A_t \left(D_r e^{-D_r S_{TL}}\right)}{F_p \left(D_r e^{-D_r S_{TL}} + 2A_t\right)}. \tag{7}$$

In OSNs, if $P(x)$ is the probability of a user x being an anomaly, then the likelihood $L(a)$ [34] for identifying it is given as:

$$L_a = \prod_{i=1}^{s}\left[\prod_{j=1}^{c}\left[\prod_{k=1}^{n}\left[(P_x)^{\eta}\,(1-P_x)^{n-\eta}\right]_k\right]_j\right]_i, \tag{8}$$

where s is the number of social networks, c is the number of communities, n is the number of users, and η is the number of users identified as potential anomaly such that $\eta \leq n$. The likelihood can be maximized to obtain the actual set of anomalous users by minimizing the similarity between the probability of a user being anomalous and probability that none of the user is anomalous. Eq.(8) can be computationally expensive, which can be resolved by the parallel implementation of the proposed solution.

The asymptote for anomaly detection in multiple OSNs depends on D_o, R_c, F_p and T_L. These parameters govern the final detection of potential anomalies in OSNs, before being fed into the outlier to accurately map the anomalous users and communities. From Eqs.(3) and (4),

$$F_{p,A_t} = F_{p,A_{t-1}}\left(1+\frac{D_o}{R_c T_L}\right)^{A_t}. \tag{9}$$

If the network is treated as a single static labeled graph, then $A_t = 1$, which implies

$$\frac{F_{p,A_t}}{F_{p,A_{t-1}}} = 1+\frac{D_o}{R_c T_L},$$
$$\frac{D_o}{T_L} = -R_c\left(\frac{F_{p,A_t}}{F_{p,A_{t-1}}}-1\right). \tag{10}$$

Eq.(10) is the asymptote for determining the anomalies in OSNs considering a single static labeled graph. However, for enhancing the detection of anomalies, it is required to consider the dynamic labeled graphs which considerably depend on A_t such that Eq.(3) deduces to

$$\frac{D_o}{T_L} = -R_c\left(\left(\frac{F_{p,A_t}}{F_{p,A_{t-1}}}\right)^{-A_t}-1\right). \tag{11}$$

From Eqs.(10) and (11), it can be noticed that for accurate determination of an anomaly in OSNs, there must be a considerable change in F_p, otherwise it becomes inefficient and difficult to find the faulty users. Thus, in order to overcome such situations, outliers are operated over the final set of potential users identified on the basis of fair play points. The methodology opted for the identification of anomalies in cross platform OSNs is shown in Fig. 4. The system model forms the basis of the proposed intelligent sensing model which is then operated over by the filtering algorithms to finalize the decision on anomalies as well as on the classification of communities.

VI. PROPOSED APPROACH

It is difficult to detect the cross-platform anomalies since these are based on the interactions between different sources

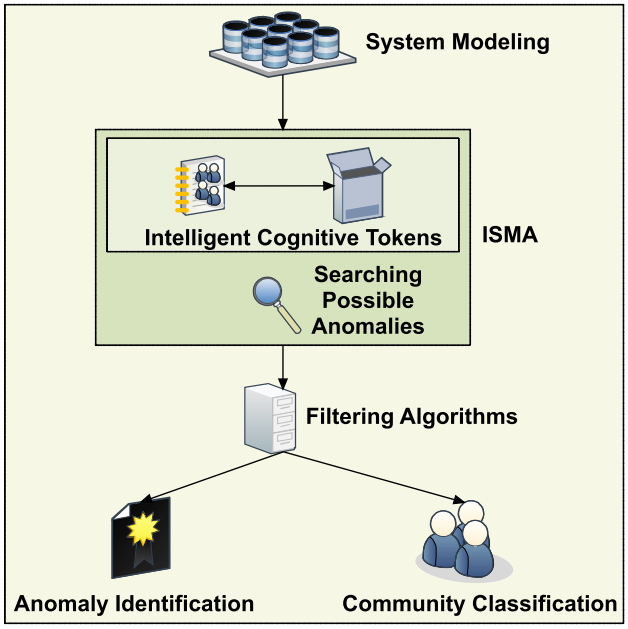

FIGURE 4. Methodology for anomaly detection.

in a different manner unlikely the vertical anomalies which are based on the type of threats caused by the users. Cross platform anomalies can be either intentional or unintentional. Thus, it is required to distinguish between the intentional as well as unintentional anomalies amongst the cross-platform anomalous users. The proposed approach utilizes the features defined in the system model and builds up a model which can detect anomalies in two phases. The first phase is the sensing platform that provides an intelligent sensing solution to the potential threats in the multiple OSNs, and the second phase is the outlier which finally identifies the anomalies and performs community classification. The two phases of the proposed approach are explained below.

A. ISMA: INTELLIGENT SENSING MODEL FOR ANOMALIES

The proposed approach aims at an intelligent solution for identifying the cross-platform anomalies across multiple OSNs. The intelligent approach is derived using the concept of "Cognitive Tokens (CTs)". CTs are the intelligent sensing data gatherers of a network which are deliberately induced in the network to attract wrong users. These are the sort of attraction links that act as a fodder for an anomalous node operating across the multiple platforms. A single OSN can have multiple CTs. These CTs can be a simple hyperlink, social ads, or any data exchange tool. CTs are capable of tracking major of the features required for the identification of a potential anomaly. In order to counterfeit the risk of identification by anomalous users, CTs are regularly updated and a log is maintained for off-site as well as on-site analyses. Off-site analyses are performed at the data centers which store the data of a particular social networking site, whereas on-site analyses are performed by the hosting server itself.

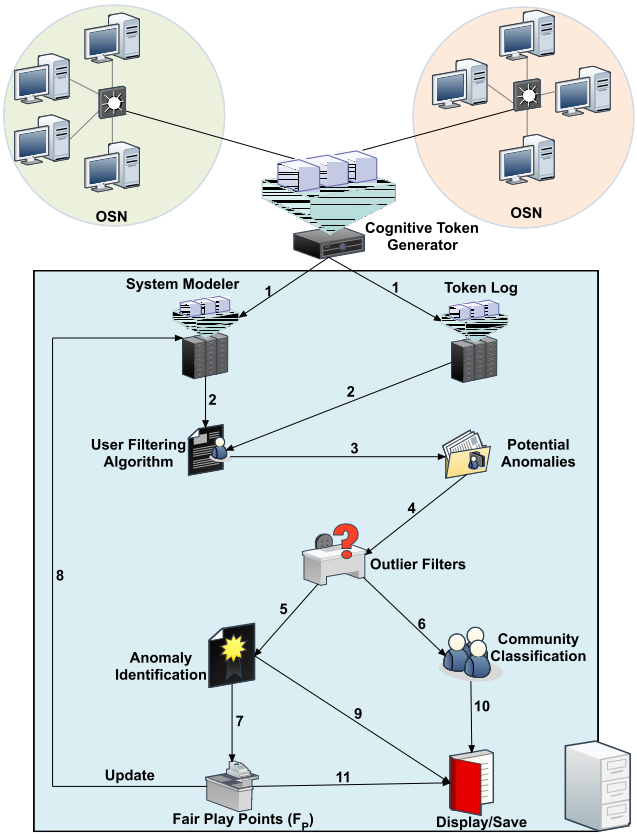

FIGURE 5. Proposed intelligent sensing model for anomaly detection in cross-platform OSNs.

Algorithm 1 ISMA: Anomaly Filtering Algorithm

Input: N, G, S_{TL}, T_L
i=1
Convert dynamic labelled graph to static graph
set S_{TL}^{TH}, T_L^{TH}
potential_anomaly[]
while $i \leq |N|$ **do**
  Calculate S_{TL} and T_L
  if $S_{TL} \geq S_{TL}^{TH}$ && $T_L \geq T_L^{TH}$ **then**
    potential_anomaly[]=i
  end if
  i=i+1
end while

Algorithm 2 Dynamic Labeled to Static Graph Conversion

Input: $G = (|N|, E, \pm F_c), A_t$
G'=$G = (|N|, E, F_c)$
k=count(features)
i=1
while $i \leq |N|$ **do**
  E[i]=sum of F_c on all edges for node i
  $normalized[i] = \frac{F_c(i) - \min(E[i])}{\max(E[i]) - \min(E[i])}$
  $W[i] = \max(F_c(i)) \times normalized[i]$
  i=i+1
end while

The detailed procedural diagram of the proposed intelligent Sensing Model for Anomalies (ISMA) is shown in Fig. 5. The steps given in the figure show two planes which coordinate the task of finding anomaly amongst OSNs users.

The first part of the architecture includes a cognitive token generator. The generator works in coordination with the server hosting a particular site or can be integrated with the data centers where analyses are to be performed. The CT generator operates as an individual entity and creates anomalous links to attract users which may or may not cause threat. Each entry generated by a CT generator is maintained as a log on a token log server. Token log server is the same server on which the entire procedure of identifying an anomaly is carried. The server forms the second part of the proposed model. All the major steps are accomplished on the server except for generating the CTs. According to the diagram, system modeler, and token logger takes into the data provided by each CT generator, which is then fed as an input to the user filtering algorithm. The user filtering algorithm distinguishes the normal users from potential anomalies across them by utilizing S_{TL} and T_L. The detailed steps for the identification of potential anomalous users are shown in Algorithm 1. The algorithm uses another procedure to convert the dynamically labeled graphs into a single static labeled graph as given in Algorithm 2. The same algorithm can be used for multiple as well as single instance graphs. The proposed approach uses F_c as a measure of weight for the formation of dynamic graphs over the interaction time A_t. Algorithms 1 and 2 operate with a runtime complexity of the order $O(n)$, where 'n' is the number of users.

The thresholds for S_{TL} and T_L are pre-decided. The decision is made on the basis of token generation rate and interaction of users. In the proposed approach, all the thresholds considered for the features presented in the system model are derived in the theoretical analyses part. The potential anomalies are then fed into the error-based outlier that finally takes a decision on the identification of anomaly, and also classifies the communities based on the user activities. Next, both these outcomes are saved on a server or can be displayed as discussed earlier to show fair play points to every user either publicly or privately. Finally, F_p is updated for every user and the system modeler and token generator continues for next iteration. The iterations can be governed by the time slot similar to time series approach. The CT generator keeps on performing its activities and an external source takes the log whenever needed. Thus, the proposed ISMA can serve the purpose of both on-site as well as off-site analyses for the detection of the anomalous user. For strict implementation, ISMA should be put on continuous monitoring state. Since there is no concept of CTs in the existing scenarios, by the medium of ISMA, it is emphasized to utilize social indexing as a tool against the anomalies across the cross-platform OSNs.

Algorithm 3 EOF: Error-Based Outlier Filter

Input: G, N, D_o, D_r, R_c, A_t
i=1
set F_p^{TH}, E_g^{TH}
final_anomaly[]
while $i \leq |N|$ **do**
    $F_{p,A_t} = F_{p,A_{t-1}} \left(1 + \frac{D_o}{R_c T_L}\right)^{A_t}$
    $E_g = \frac{2A_t(D_r e^{-D_r S_{TL}})}{F_p(D_r e^{-D_r S_{TL}} + 2A_t)}$
    if $F_p \leq F_p^{TH}$ && $E_g \geq E_g^{TH}$ **then**
        final_anomaly[]=i
    end if
    i=i+1
end while
set community_thresholds=[]
set k=count(community_thresholds=[])
community[]
j=1
while $j \leq k$ **do**
    H=community_thresholds=[j]
    p=1
    while $p \leq |N|$ **do**
        if $F_p \leq H$ **then**
            community[p]=j
        end if
        p=p+1
    end while
    j=j+1
end while

B. EOF: ERROR-BASED OUTLIER FILTER

Outlier filters are the simple tools for determining the user community and filtering the anomalies. Filters can be density-based or distance-based filters. Density-based filters are used whenever there is a requirement of finding an anomaly depending on the behavior w.r.t. other members of the community. The existing filters operate on the similar principle, thus, cannot be readily applied with the ISMA. Thus, a novel error-based outlier filtering (EOF) is proposed which operates in coordination with the ISMA to finalize the anomaly along with the identification of influential as well as the influenced communities. It is to be noted that the existing outlier algorithms have to go through each user of OSNs to distinguish them into communities which then decides the possible anomaly in the network, whereas the EOF will look for only those users which fall under the category of potential anomalies. The search amongst the lesser number of users with more appropriate features of OSNs allows efficient classification and identification of the anomalous users. The EOF approach operates over two major features, namely E_g and F_p, which are governed by the threshold similar to the anomaly detection algorithm. The steps for determining final anomaly and the community classification are presented in Algorithm 3. The runtime complexity of this algorithm is of the order $O(n\ \ k)$, where k is the number of community divisions to be made in the entire social networks. The EOF plays a crucial role in the proposed ISMA for identification of anomalies. A parallel run of EOF for multiple OSNs supports the algorithm in finding results with lesser search-time complexity. The same server hosting the ISMA is used for EOF. The EOF algorithm uses a threshold mechanism to classify the users into different communities. This threshold can be divided into the least possible entities to categorize communities into smaller possible groups. However, the limiting value of communities up to which users can be classified is obtained trivially and is always less than equal to $\frac{1}{F_p^{TH}} \sum^{|N|} F_p$.

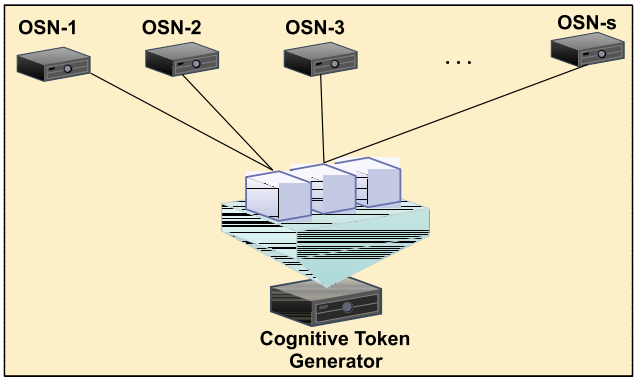

FIGURE 6. Cognitive token generator dependency on different OSN servers.

C. COMMON VERIFICATION PLATFORM FOR MULTIPLE OSNs

ISMA depends on the extra facility of CT generator required to be deployed by the OSN service providers. This adds up the cost of maintenance for the service providers. However, from the fact of lowering the risk of cross-platform anomalies, the operability cost is bearable. But, sill keeping this in view, an alternative solution to the problem is identified which is termed as the CT generator sharing. Initially, each of the controllers of social networking sites has to maintain a separate generator for CTs, which are then accumulated at the CT logger at the respective data analyses centers of each OSN. This is a complex and an expensive task in itself, as shown in Fig. 6. However, with a single login mechanism for every OSN, the cost can be decreased, and the task of identifying anomalous users across the multiple OSNs can be improved. A simple approach to common login mechanism and utility of single CT generator is shown in Fig. 7. CT generator is installed as a third party with access to APIs of all the major OSNs. The common CT generators are operated in coordination with the common login servers, which utilize a unique identification model for every user such as ''National ID'', ''Mobile No.'', etc. With the advent of Apps-based internet, it has become easy to utilize the features of common login. Thus, this approach is very much practical and can be considered as a new business model for the upcoming market of online social media as well as device dependent IoT.

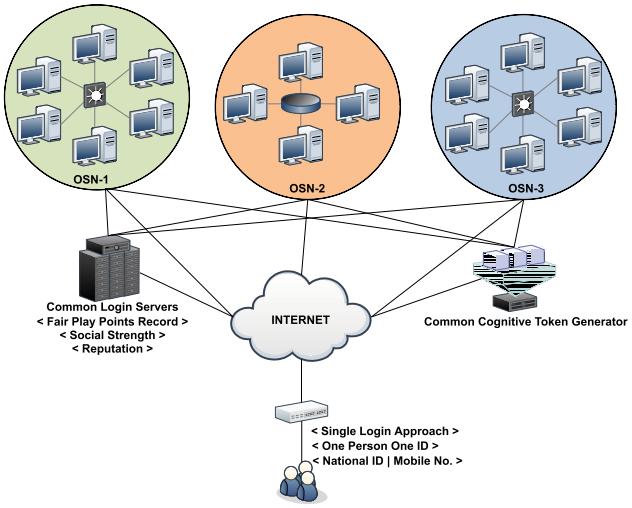

FIGURE 7. Common Verification Platform for Multiple OSNs.

VII. THEORETICAL ANALYSES

This section presents the theoretical evaluation of the proposed approach and determines the thresholds to be set for the detection of anomalies across the cross-platform OSNs.

Lemma 1: With the increase in S_{TL} to a larger value, the detection of anomaly is difficult since error in fair-play is unidentifiable.

Proof: From Eq.(7),

$$E_g \propto e^{-S_{TL}}, \tag{12}$$

which means that with S_{TL} attaining a maximum value, $E_g = 0$. Thus, it becomes difficult to utilize outlier for an extremely higher value of S_{TL}. This higher value means that there are number of alternative routes between the anomalous and non-anomalous users, which is difficult to trace and identify.

Lemma 2: With lesser number of alternative paths, the probability of detection of an anomaly increases, which increases the overall likelihood of the system for anomaly detection.

Proof: From Eqs.(10) and (11),

$$\frac{D_o}{T_L} \propto -R_c, \tag{13}$$

which implies a decrease in the detection asymptote with an increase in the R_c. The result shows that with more number of alternative paths available for the anomalous user to interact with other users, the relation cost increases, which decreases the probability of detecting an anomaly. The reverse also holds in this situation, and a decrease in the number of alternative paths increases the asymptote limit of the OSNs for detecting anomalies, which increases the probability of identification of anomalies resulting in the increase of overall likelihood for anomaly detection (ref. Eq. 8).

Remark 1: With the consistent increase in the social strength (D_o), the fair play points F_p increase resulting into the decrease in E_g. F_p can be easily increased by using Sybil attack. Thus, in order to prevent Sybil attack in the cross-platform OSNs, the difference in F_p w.r.t. previous value should be monitored consistently.

Proof: The continuous increase in D_o, causes large scale-up in the value of F_p as compared to previous value as seen in Eq.(2). The value shows that if fair-play points of the user increases, it becomes difficult to track and identify such anomalous user. Such situation is the major consequences of a Sybil attack. Thus, Sybil attacks must be prevented in cross-platform OSNs.

Remark 2: The thresholds for S_{TL}, T_L are accounted considering the unintentional and intentional errors. Thus, selecting threshold entirely depends on the number of users to be considered as a potential anomaly. In general, the threshold for the proposed approach is calculated as the average of the sum of the mean and minimum value for the users under evaluation.

Proof: The selection of the threshold values are hypothetical and can vary from case to case. However, a lower value allows easy and efficient detection of anomalies. A threshold with lower value allows enough error in fair play (Ref. Eq.7) and also F_p (Ref. Eq. 4), which makes it easier to detect most of the anomalies. However, for strict evolutions, the threshold for E_g should be taken minimum from all values.

VIII. PERFORMANCE EVALUATION

The proposed ISMA approach is formed on the basis of a futuristic ideology of incorporating suggested features for every user in social networks. Also, the creation of single point entry can decrease the complexity associated with the data storage and processing. The proposed approach depends on two algorithms, the anomaly detection, and error-based outlier filtering. Since none of the existing OSN platforms utilize the concept of CTs as proposed in this paper, a synthetic data is used to check the effectiveness of the proposed solution. The synthetic data analyses are carried using *Matlab*^{TM} with visualization support from *Gephi*^{TM} libraries. However, in the later part, a real-time email dataset is also used to evaluate the performance of the proposed outlier approach in comparison with the existing solutions.

A. SIMULATIONS AND RESULTS OVER SYNTHETIC DATA SET

Three OSNs are created similar to the one presented in the proposed solution. The number of users in the social networks varied as 450, 300, 250, respectively, as shown in Fig. 8. Total nodes used for analyses are 1000 with 34510 edges each having a different value for F_c. The average degree of the entire network is 12 with an average distance between the source and target of 5 and minimum distance 2. The minimum connected components are taken 2 in the dataset. In the synthetic data, the routers and switches are not considered for any traffic evaluations. However, a layered model is used similarly to the one shown in ISMA. The thresholds for all the features during simulations are derived using mean-value concept. The simulations can be easily conducted for the

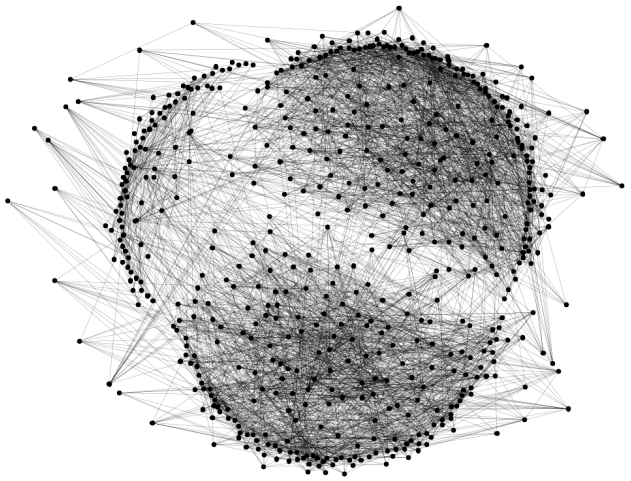

FIGURE 8. An illustration of synthetic dataset comprising 1000 users from three OSNs.

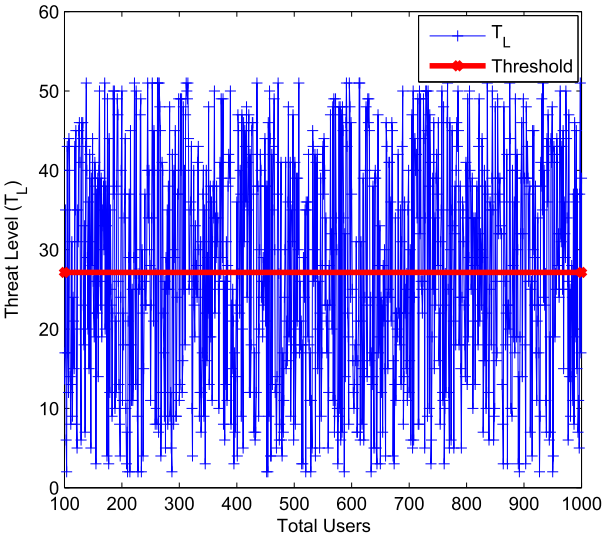

FIGURE 10. Threat level (T_L) for 1000 users.

variations in thresholds. However, the approach will remain the same, because of which results with variations in thresholds are not presented in this paper. The datasets provided can be easily used to check the variation of the proposed model in contrast to variation in the thresholds for different features.

depending on the interaction of the users with the CTs. The potential anomalies are then operated over by the proposed outlier Algorithm 2, which classifies the final anomalies in the OSNs.

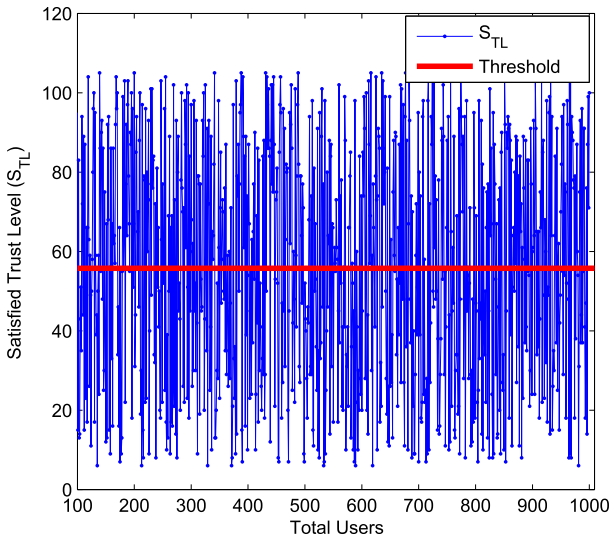

FIGURE 9. Satisfied trust level (S_{TL}) for 1000 users.

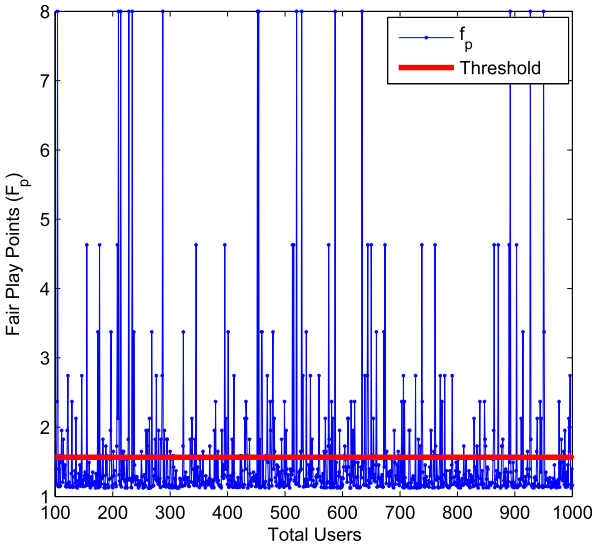

FIGURE 11. Fair play points (F_p) for 1000 users.

In simulations, 100 CTs are floated in the OSNs and the maximum value recorded for T_L is 50. With each user making an interaction with the floated CT, the values for S_{TL} and T_L are logged by the token logger. The average results for these metrics during simulations are shown in Figs. 9 and 10. The results show S_{TL} and T_L distributions for users. The thresholds for both the features are marked by a red line which is 57 for S_{TL} and 28 for T_L. Both the entities helped in the identification of potential anomalies using Algorithm 1. The variation in the hits to CTs decides the value for F_p for each user. This value may decrease or increase in the next iteration,

However, the simulation traces show that the users interested in interaction with CTs are mostly anomalous and do not restrain from further interactions. The variation in F_p for different users and its threshold at 1.7 is shown in Fig. 11. This value is used to identify E_g, which forms the basis for the formation of the proposed outlier algorithm. The value of E_g and its threshold at 4.1 is shown in Fig. 12. The values plotted in the graph identify the potential anomalies. In the simulations conducted for the evaluation of ISMA, the final anomalies are again checked for F_p and if any value is misinterpreted as an anomaly, it is removed from the anomalous data. The final

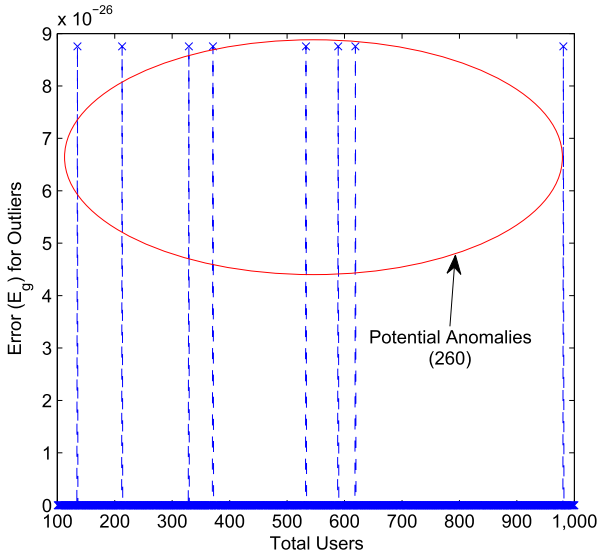

FIGURE 12. Error in fair-play E_g for 1000 users.

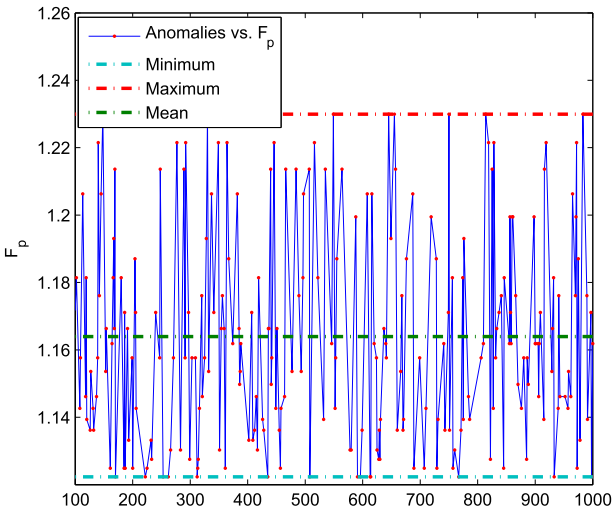

FIGURE 13. Detected anomalies and their variation in F_p.

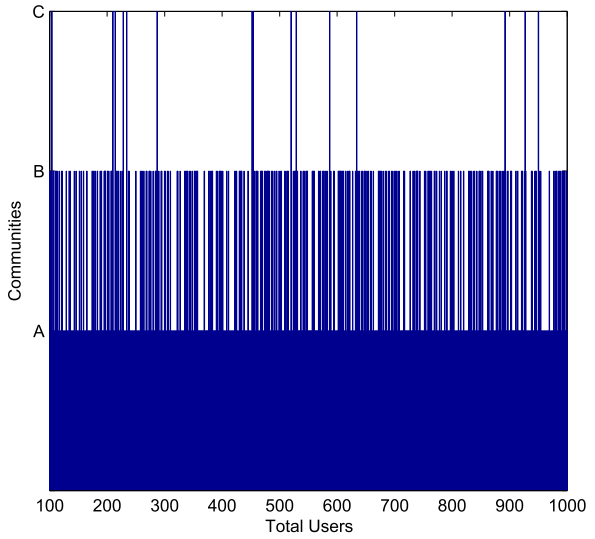

FIGURE 14. Community classification of all the users.

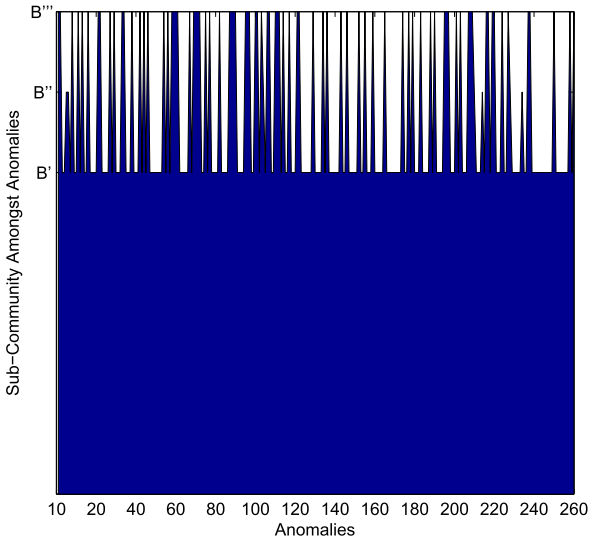

FIGURE 15. Sub-community classification for anomalous users.

plot for variation in the F_p for the anomalies identified using the proposed approach is shown in Fig. 13.

In the next part of simulations, the proposed ISMA classifies the users for communities to understand the impact of anomalous users. Not only individual users are classified, but the anomalies are also identified to understand the level of threat caused by them. Figs. 14 and 15 show the community classification into three sets for the users which are initially considered in the evaluation. The three communities, namely A, B, C are assigned names alphabetically. The safe users are presented in the community A, with B having anomalies and C with potential anomalies for OSNs. The anomalies in community B are further classified on the basis of their values for F_p and sub-categories are created in the main community to understand the impact of each anomaly.

As a broader view, the initial graph available over the course of simulations is dynamically labeled, which is converted into a single static labeled graph using Algorithm 3, as shown in Fig. 16. The static graph is operated over by the proposed approach, which separates the non-anomalous and anomalous users as shown in Fig. 17 and Fig.18, respectively.

B. STATISTICAL ANALYSES

The proposed approach is evaluated for its statistical behavior in the detection of anomalies using synthetic data. The probability distribution tool is used to statistically validate the proposed ISMA detection approach. The statistical analyses confirm the performance of the proposed approach over the considered dataset. The analyses presented in Fig. 19 suggest the users' classification and the probability of users for being in a particular community. It depicts the majority in user distribution, whereas Fig. 20 presents the possible classifications which is done over the varying value of fair play points (F_p).

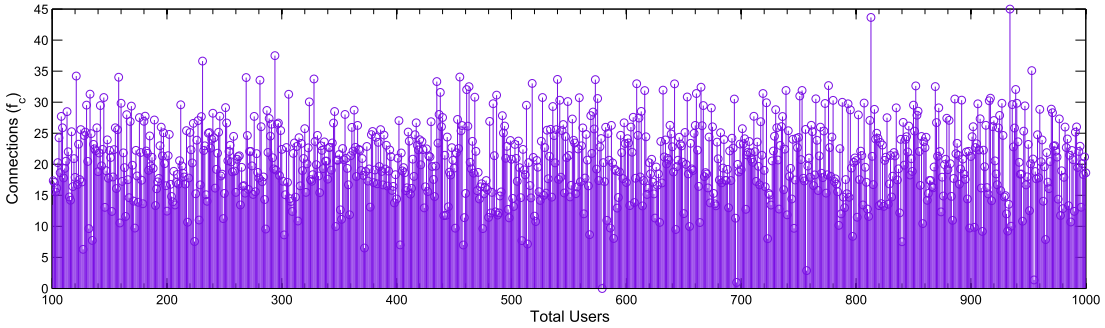

FIGURE 16. Single labeled static graph weight variation generated from multiple dynamic graphs for F_c.

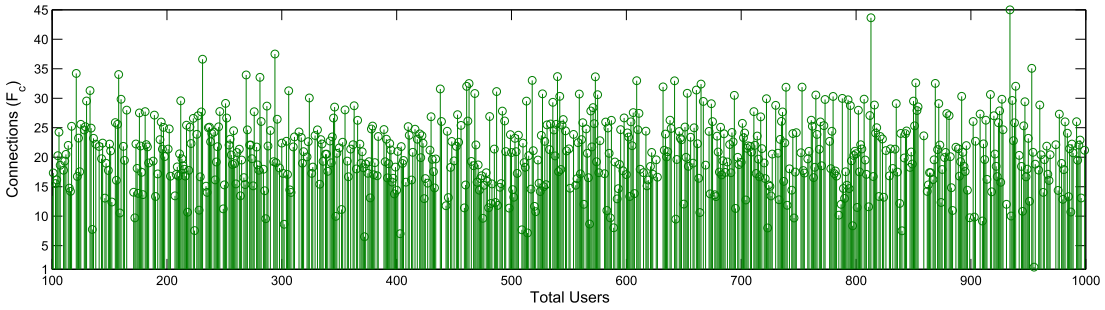

FIGURE 17. Non-anomalous users vs. F_c variations.

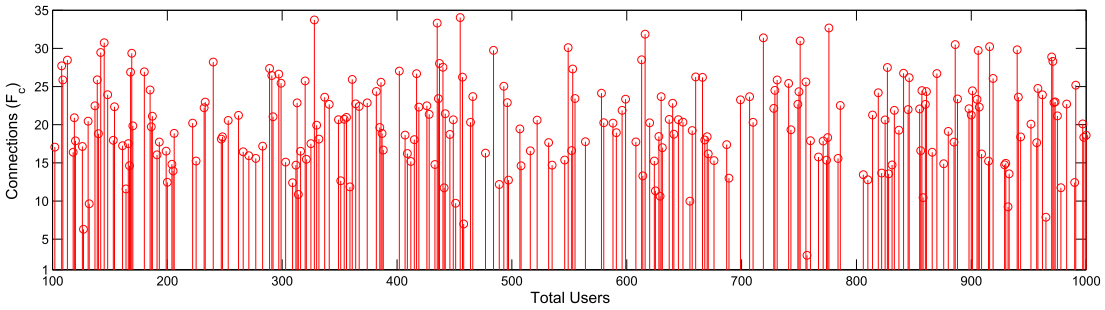

FIGURE 18. Anomalous users vs. F_c variations.

The entire approach depends on the cognitive tokens which are used to track and identify the anomalous users, thus, the density of user interaction w.r.t. S_{TL} is tracked, as shown in Fig. 21. The figure shows the level of interaction made with the false sources which are deliberately created for the identification of anomalies. Finally, the initial and end results are statistically analyzed for variation in the density w.r.t. F_c as shown in Fig. 22. The graphs present the static labeled graph weights distribution during the initial phase and after the detection of anomalies. The distributions for non-anomalous as well as anomalous users are shown as separate trend lines in the figure. These results statistically classify and demonstrate the efficiency of proposed approach in identification and classification of the anomalous and non-anomalous users.

IX. STATE-OF-THE-ART COMPARISONS

The proposed approach depends on various new features for the detection of anomalies in the cross-platform OSNs. There is no such approach which utilizes the similar features to extract anomalies out of the multiple OSNs. However, to validate and prove the effectiveness of the proposed approach, ISMA is compared with two major detection and classification approaches, namely, Support Vector Machine (SVM) [11], and k-nearest neighbor classifier [10]. The email dataset is used for the evaluation of ISMA, SVM, and k-nearest neighbor approaches. Previous one-year email data is gathered from authors email account (provided as supplementary files), which comprises 6000 entries from 64 unique interactions. The dataset contains links between the receivers and senders. The data is gathered by labeling each email with a unique ID, and separate IDs are given to the sender and receiver of the emails. Finally, all are grouped together to generate the email dataset. For anonymity, the details are provided only in the form of numeric data which represents a number of interactions. The entire data is categorized into 10 dynamic graphs merged together using the policies explained in the Section VI.

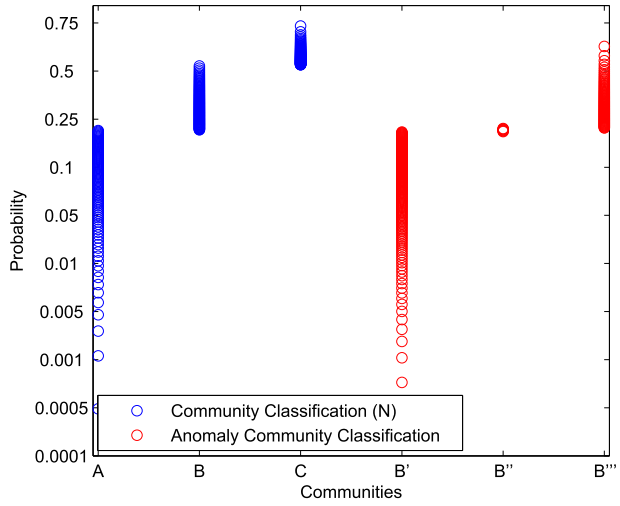

FIGURE 19. Probability of user distribution vs. communities.

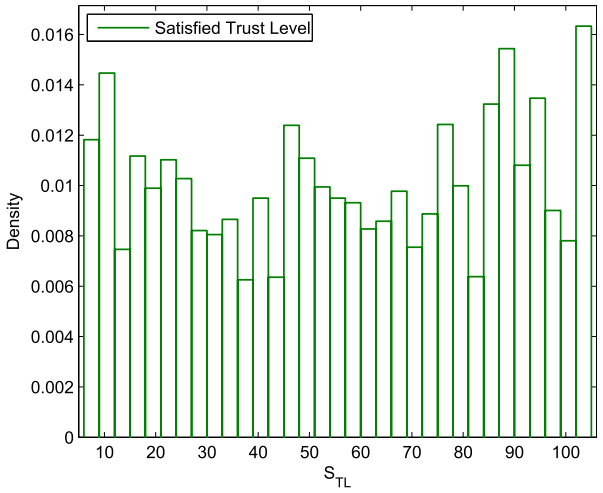

FIGURE 21. Density variation in S_{TL} for ISMA over synthetic dataset.

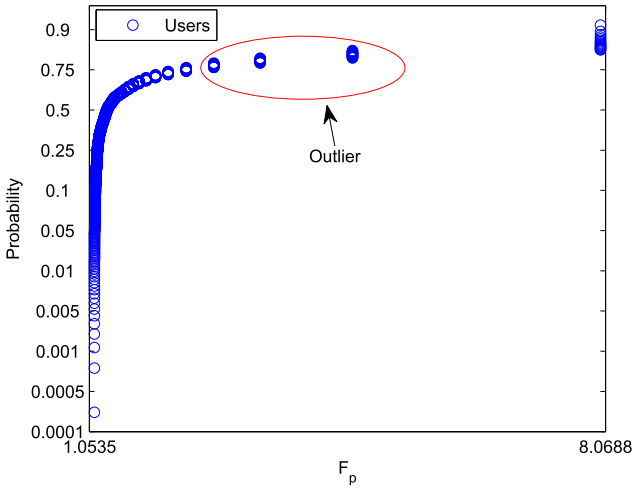

FIGURE 20. Probability of user distribution vs. F_p.

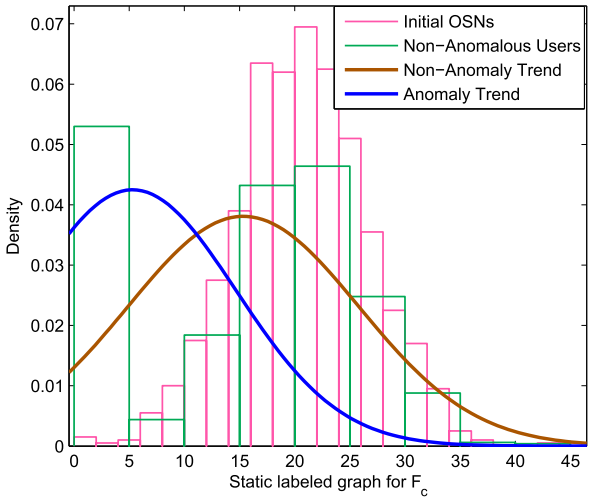

FIGURE 22. End results and graph trends with variation in F_c.

Each number in the dataset represents a source with which an email interaction is made. Each value represents a unique ID given to an email interaction depending on the source. Here, sources include other users, spam emails, advertisements as well as update emails. For 10 dynamic links, the most interactions are made with the users #35, #39, #30, #62, #25, #50, #21, #4, #6, and #49. For each of these ten graphs, a common output anomaly is chosen which acts as the target assigned to the considered approaches for comparative analyses. The output is the spam emails which are to be detected on the basis of links and emails exchanged between the sender and different receivers. The output file contains values in the form of 0 and 1, where 1 represents an anomaly and 0 represents non-anomalous data exchange. The detailed datasets are provided as the supplementary files. The considered approaches are evaluated for accuracy in identifying the anomalies and error range in the identification of anomalies.

The proposed approach is evaluated against the existing SVM approaches, using Radial Basis Function (RBF) and sigmoid functions over the cost of classification C and γ of Gaussian Kernel, and the k-nearest neighbor approach with distance equal to 2 between the inputs and outputs as shown in Table 2. The proposed approach is able to provide the highest accuracy of 99.2% which is 25.1% higher than the SVM-RBF and sigmoid approach, and 22% higher than the k-nearest neighbor approach. Further, the proposed ISMA also cause less error in detecting the anomalies which are within the range of 0.1% to 2.8%. On the best operations, the error is 96.6% lesser than the SVM and k-nearest neighbor approaches. The comparative results presented in this section validate the efficiency of the ISMA in identification and classification of anomalies in cross-platform OSNs.

X. A CASE STUDY ON ANOMALY DETECTION IN IoT

IoT is one of the important examples of cross platforms systems which may suffer from horizontal threats. Different users connect to different devices through multiple gateways [13]. Each user may act as a positive user for one device

TABLE 2. State-of-the-art comparison and results.

Approach	Parameter	Accuracy	Error
SVM-Radial Basis Function (RBF)	C, γ	74.21% (2 Fold), **74.23% (3 Fold)**, 73.34% (4 Fold), 74.21% (5 Fold)	$\pm 3\%$
SVM-Sigmoid	C, γ	73.21% (2 fold), **74.22% (3 fold)**, 74.21% (4 fold), 73.71% (5 fold)	$\pm 3\%$
K-nearest Neighbor	$X_i, X_j, 1 \times 1$	73.4% - **77.2%**	$\pm 3\%$
ISMA	$G(N, E, \pm F_c)$	94.1% to **99.2%**	$\pm$ (0.10% - 2.8%)

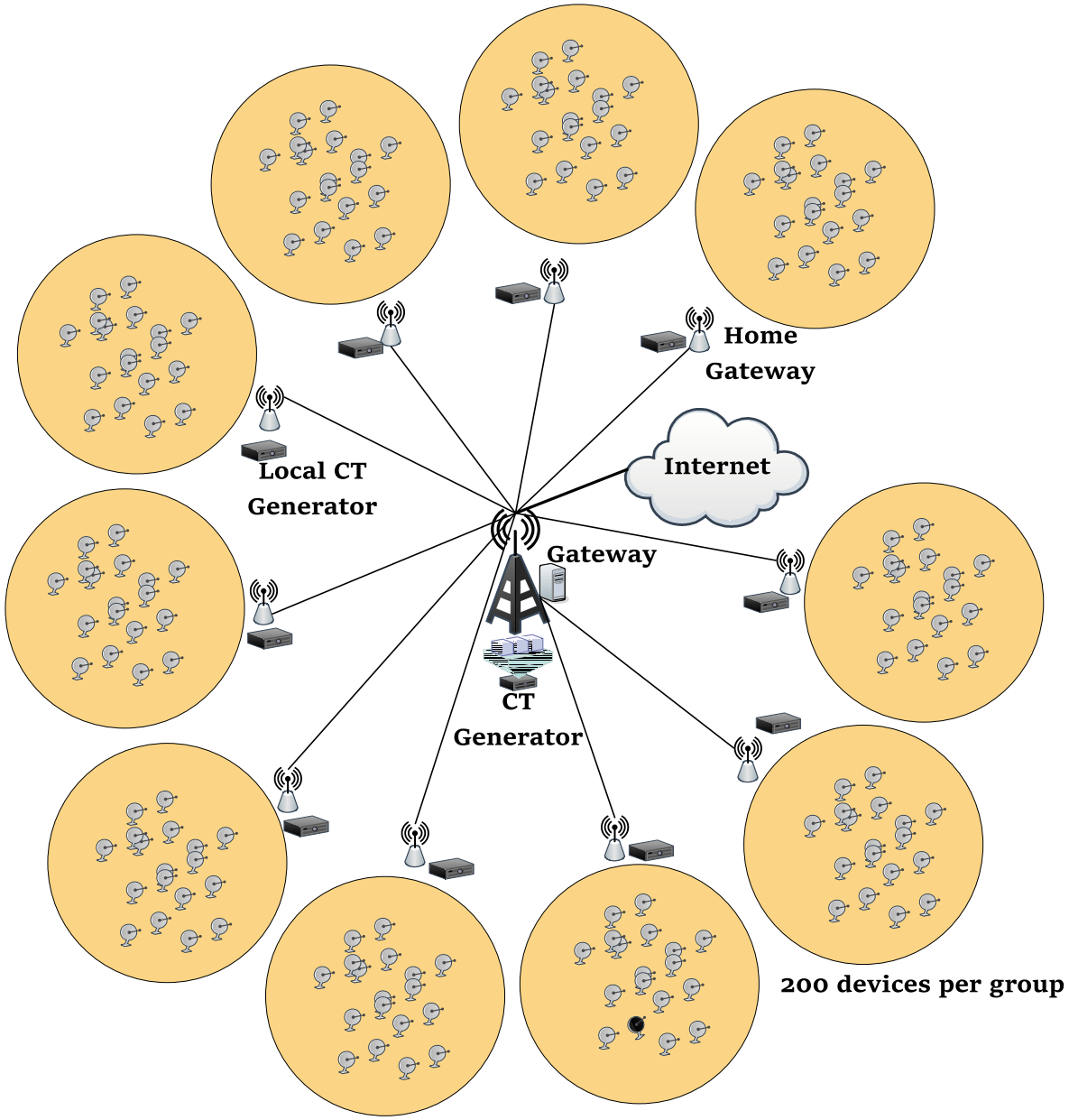

FIGURE 23. An illustration of IoT scenario for evaluation of ISMA.

and a negative user for every other device [12]. IoT devices operate in layers following a device to device as well as a device to infrastructure communications which put them under the threat of anomalies that may or may not reside in their subgroup [14], [35]. Spam filtering and remote access control are the key challenges for IoT devices [36], [37].

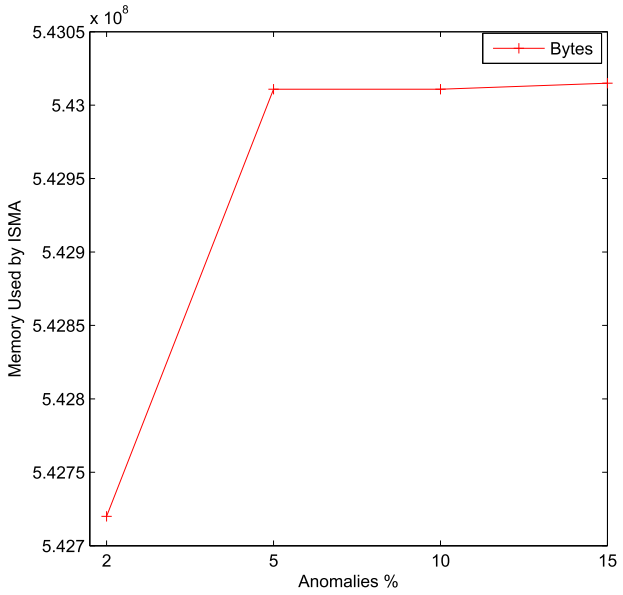

FIGURE 24. Memory utilization vs. anomalies.

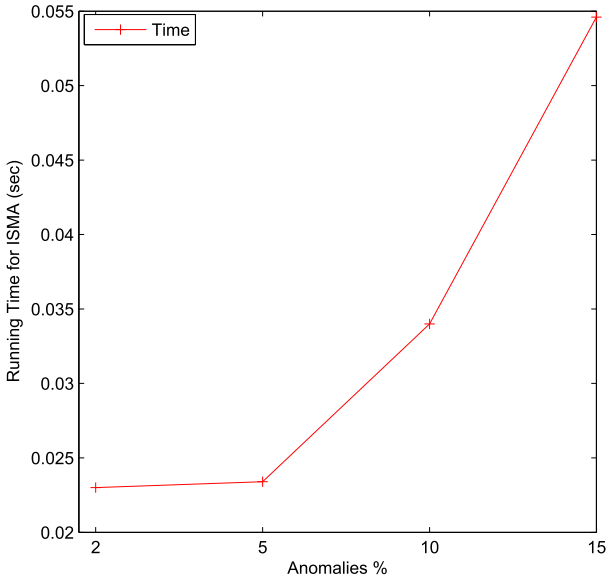

FIGURE 25. Running time vs. anomalies.

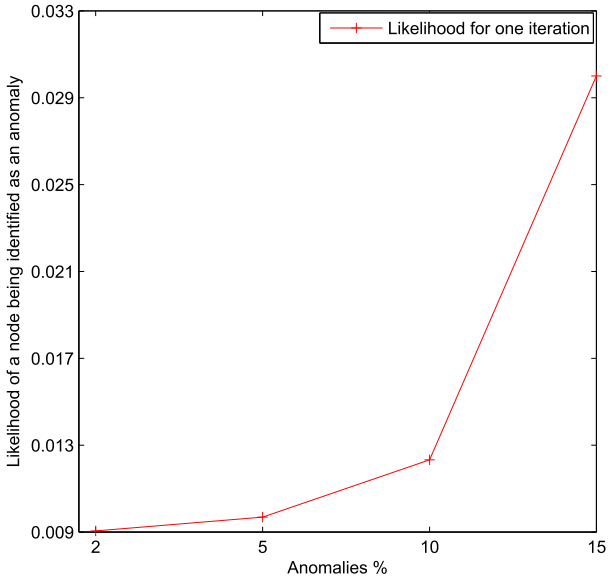

FIGURE 26. Likelihood for single iteration vs. anomalies.

The proposed approach can be used in the detection of anomalies within the same as well as multiple subgroups. In the case of IoT, CTs are the physical devices that act as a non-service provider collaborative systems, which can respond to a query, and can maintain the log on the basis of similar algorithms provided in Section VI. This case study includes simulation-based analyses of the proposed approach over the IoT.

A simple IoT environment is considered comprising 2000 devices across 10 subgroups. Each subgroup used its own CT generator which works in collaboration with the centralized CT generator of the gateway as shown in Fig. 23. All the computations and logging are done at the gateway. Local computations and logging can also be performed, but this consumes a lot of processing power as well increases overheads. Anomalies are introduced at the rate of 2%, 5%, 10% and 15%. The proposed solution is used to evaluate these anomalies with 10 CTs per subgroup. This study is independent of the communication aspects of devices and is performed in an environment without any loss. Most of the approaches make it difficult to identify anomalous users if the anomaly rate is increased. However, in the case of ISMA, with more number of potential anomalies in the network, the values of each feature as described in the system modeling enhances, which makes it easier to identify anomalous users. Since the proposed approach utilizes the concept of community classification, it can easily distinguish between the hazardous and non-hazardous nodes in IoT environment. This can be justified considering Eq.(8). The likelihood of detecting an anomaly increases with an increase in the probability of connections made by the anomalous node. The proposed approach is evaluated for the time consumed in the identification of anomalies, memory utilization, and the likelihood w.r.t. variation in the anomaly rate.

The results are traced for 50 runs with variation in the placement of the devices in the IoT environment. Although the number of devices has less impact than the number of connections, still this variation is used to identify the performance of the proposed approach in a variable topology. Figs. 24, 25, and 26 present the average values obtained for memory utilization, running time and likelihood in a single iteration. The results show that the memory consumption increases with an increase in the number of anomalies as more processing is performed. However, the increase is stabilized after 5% increase in the anomalies and do not increase linearly as it increased between 2% to 5%. Thus, the lesser increase in the memory makes ISMA suitable for large-scale networks. The time complexity of the ISMA is

directly related to the number of anomalies it has to identify. Irrespective to the memory utilization, time complexity increases exponentially which is a concern; but the time complexity can be considerably reduced by the parallel run of the proposed algorithms.

Further, the less consumption of memory and an increase in the computational time increases the likelihood of determining an anomalous node. Fig. 26 shows that an increase in the number of anomalies has more likelihood of being detected. This is because of higher S_{TL} value which increases with the increase in the number of anomalies. However, a very large value of S_{TL} makes it difficult to identify an anomaly as an anomalous user will have multiple routes of contact with different sources. Apart from ISMA being utilized for anomaly detection in IoT environments, the common login solution for all users can be a much safer and an inexpensive way of determining the anomalies in terms of computational time and memory utilization.

XI. OPEN ISSUES

Anomaly detection has always been a subject of concern in almost every research field. Concerning OSNs, anomalies identification and community classification are the most challenging issues that require efficient strategies to prevent data-counterfeits. User ranking and common public identification can be two of the efficient solutions for determining the anomalies. However, such solutions require collaboration between the service providers as well as the hosted sites. Some of the crucial issues which are yet to be tackled in the OSNs are:

- Prevention of Sybil attack on cross platforms OSNs. With a thrust to false reputation, anomalies can prove hazardous to users across different social network platforms [1].
- Prevention of fake data seeding in cross-platform OSNs.
- Classification of users under the threat of horizontal anomalies.
- Real-time services to provide anomaly detection with continuous user monitoring rather than depending on data centers for off-site evaluations.
- Formation of intelligent visualization platforms which can represent information related to anomalous nodes [38].

XII. CONCLUSION

Anomalies are a major issue for multiple OSNs. Anomalies across different platforms can compromise user information which can be hazardous to the entire community. In this paper, a problem of identifying anomalies across cross-platform OSNs was considered and an intelligent sensing model for anomaly (ISMA) detection was proposed. The proposed approach utilized the concept of cognitive tokens to identify the potential anomalies which were operated over by the proposed error-based outlier algorithm that confirms whether a user is an anomaly or not. A fair play point approach was used for the determination of the anomalies. The initial evaluation of the proposed approach was carried using synthetic dataset followed by theoretical and statistical results. Further, the proposed approach was compared with the SVM-based RBF and sigmoid approach, and the k-nearest neighbor approach. A real email dataset was considered for the state-of-the-art comparison. In order to validate the proposed approach for its generic implementation over cross platforms, a case study was presented for anomaly detection in IoT environment. The results show that the proposed approach can efficiently identify and track anomalies in IoT environments with less memory utilization and with a high likelihood of detection. The proposed approach was able to provide the highest accuracy of 99.2% which was 25.1% higher than the SVM-RBF and sigmoid approach, and 22% higher than the k-nearest neighbor approach. Further, the proposed ISMA also caused less error in detecting the anomalies which were within a range of 0.1% to 2.8%. In the best case, the error was 96.6% lesser than the SVM and k-nearest neighbor approaches. The gains in comparative results validate the efficiency of the ISMA in identification and classification of anomalies in cross-platform OSNs.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

SUPPLEMENTARY FILES

The datasets and results are provided as separate files.

REFERENCES

[1] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, "Anomaly detection in online social networks," *Social Netw.*, vol. 39, pp. 62–70, Oct. 2014.
[2] J. R. C. Nurse, A. Erola, M. Goldsmith, and S. Creese, "Investigating the leakage of sensitive personal and organisational information in email headers," *J. Internet Services Inf. Secur. (JISIS)*, vol. 5, no. 1, pp. 70–84, Feb. 2015.
[3] P. Bindu and P. S. Thilagam, "Mining social networks for anomalies: Methods and challenges," *J. Netw. Comput. Appl.*, vol. 68, pp. 213–229, Jun. 2016.
[4] C. Lu, J. X. Yu, R.-H. Li, and H. Wei, "Exploring hierarchies in online social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 2086–2100, Aug. 2016.
[5] M. G. Vigliotti and C. Hankin, "Discovery of anomalous behaviour in temporal networks," *Social Netw.*, vol. 41, pp. 18–25, May 2015.
[6] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *ACM SIGMOD Rec.*, vol. 42, no. 2, pp. 17–28, May 2013.
[7] A. Rezaei, Z. M. Kasirun, V. A. Rohani, and T. Khodadadi, "Anomaly detection in online social networks using structure-based technique," in *Proc. 8th Int. Conf. Internet Technol. Secur. Trans. (ICITST)*, Dec. 2013, pp. 619–622.
[8] A. Soule, K. Salamatian, and N. Taft, "Combining filtering and statistical methods for anomaly detection," in *Proc. 5th ACM SIGCOMM Conf. Internet Meas.*, 2005, p. 31.
[9] P. Chhabra, C. Scott, E. D. Kolaczyk, and M. Crovella, "Distributed spatial anomaly detection," in *Proc. IEEE 27th Conf. Comput. Commun. (INFOCOM)*, pp. 1705–1713, Apr. 2008.
[10] Y. Liao and V. R. Vemuri, "Use of K-nearest neighbor classifier for intrusion detection," *Comput. Secur.*, vol. 21, no. 5, pp. 439–448, Oct. 2002.
[11] S. Theodoridis, A. Pikrakis, K. Koutroumbas, and D. Cavouras, *Introduction to Pattern Recognition: A MATLAB Approach*. San Diego, CA, USA: Academic, 2010.
[12] S. Raza, L. Wallgren, and T. Voigt, "SVELTE: Real-time intrusion detection in the Internet of Things," *Ad Hoc Netw.*, vol. 11, no. 8, pp. 2661–2674, Nov. 2013.

[13] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Generat. Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660, 2013.
[14] S. Bin, L. Yuan, and W. Xiaoyi, "Research on data mining models for the Internet of Things," in *Proc. Int. Conf. Image Anal. Signal Process. (IASP)*, Apr. 2010, pp. 127–132.
[15] N. Tsitsiroudi, P. Sarigiannidis, E. Karapistoli, and A. A. Economides, "EyeSim: A mobile application for visual-assisted wormhole attack detection in IoT-enabled WSNs," in *Proc. 9th IFIP Wireless Mobile Netw. Conf. (WMNC)*, Jul. 2016, pp. 103–109.
[16] F. Baiardi, F. Tonelli, A. Bertolini, and R. Bertolotti, "Selecting countermeasures for ict systems before they are attacked," *J. Wireless Mobile Netw., Ubiquitous Comput., Dependable Appl.*, vol. 6, pp. 58–77, Jun. 2015.
[17] H. H. Pajouh, R. Javidan, R. Khayami, D. Ali, K.-K. R. Choo, "A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in IoT backbone networks," *IEEE Trans. Emerg. Topics Comput.*, doi: 10.1109/TETC.2016.2633228, 2016.
[18] A.-M. Abeer, H. Maha, A.-S. Nada, and M. Hemalatha, "Security issues in social networking sites," *Int. J. Appl. Eng. Res.*, vol. 11, no. 12, pp. 7672–7675, 2016.
[19] N. Kökciyan and P. Yolum, "PriGuard: A semantic approach to detect privacy violations in online social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 10, pp. 2724–2737, Oct. 2016.
[20] Z. Tan, J. Ning, Y. Liu, X. Wang, G. Yang, and W. Yang, "ECRModel: An elastic collision-based rumor-propagation model in online social networks," *IEEE Access*, vol. 4, pp. 6105–6120, 2016.
[21] B. Baingana and G. B. Giannakis, "Joint community and anomaly tracking in dynamic networks," *IEEE Trans. Signal Process.*, vol. 64, no. 8, pp. 2013–2025, Apr. 2016.
[22] X. Ruan, Z. Wu, H. Wang, and S. Jajodia, "Profiling online social behaviors for compromised account detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 11, no. 1, pp. 176–187, Jan. 2016.
[23] Y. Chen, S. Nyemba, and B. Malin, "Detecting anomalous insiders in collaborative information systems," *IEEE Trans. Depend. Sec. Comput.*, vol. 9, no. 3, pp. 332–344, May/Jun. 2012.
[24] L. Liu and H. Jia, "Trust evaluation via large-scale complex service-oriented online social networks," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 11, pp. 1402–1412, Nov. 2015.
[25] Y. Liu, Y. Sun, S. Liu, and A. C. Kot, "Securing online reputation systems through trust modeling and temporal analysis," *IEEE Trans. Inf. Forensics Secur.*, vol. 8, no. 6, pp. 936–948, Jun. 2013.
[26] A. Thapa, M. Li, S. Salinas, and P. Li, "Asymmetric social proximity based private matching protocols for online social networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 6, pp. 1547–1559, Jun. 2015.
[27] S. Gregory, "Finding overlapping communities in networks by label propagation," *New J. Phys.*, vol. 12, no. 10, p. 103018, 2010.
[28] W. Jiang, J. Wu, F. Li, G. Wang, and H. Zheng, "Trust evaluation in online social networks using generalized network flow," *IEEE Trans. Comput.*, vol. 65, no. 3, pp. 952–963, Mar. 2016.
[29] L. Cao, Y. Ou, and P. S. Yu, "Coupled behavior analysis with applications," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 8, pp. 1378–1392, Aug. 2012.
[30] A. Thapa, W. Liao, M. Li, P. Li, and J. Sun, "SPA: A secure and private auction framework for decentralized online social networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 8, pp. 2394–2407, Aug. 2016.
[31] S. Asur, S. Parthasarathy, and D. Ucar, "An event-based framework for characterizing the evolutionary behavior of interaction graphs," *ACM Trans. Knowl. Discovery Data*, vol. 3, no. 4, Nov. 2009, Art. no. 16.
[32] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. E. Zhao, "Analyzing patterns of user content generation in online social networks," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jun./Jul. 2009, pp. 369–378.
[33] J. G. W. Raaijmakers, "Statistical analysis of the michaelis-menten equation," *Biometrics*, vol. 43, no. 4, pp. 793–803, Dec. 1987.
[34] V. Sharma, H.-C. Chen, and R. Kumar, "Driver behaviour detection and vehicle rating using multi-uav coordinated vehicular networks," *J. Comput. Syst. Sci.*, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.jcss.2016.10.003
[35] B. Arrington, L. Barnett, R. Rufus, and A. Esterline, "Behavioral modeling intrusion detection system (BMIDS) using Internet of Things (IoT) behavior-based anomaly detection via immunity-inspired algorithms," in *Proc. 25th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Aug. 2016, pp. 1–6.
[36] A. Skovoroda and D. Gamayunov, "Securing mobile devices: Malware mitigation methods," *J. Wireless Mobile Netw., Ubiquitous Comput., Dependable Appl.*, vol. 6, no. 2, pp. 78–97, Jun. 2015.
[37] J. Ninglekhu, R. Krishnan, E. John, and M. Panday, "Securing implantable cardioverter defibrillators using smartphones," *J. Internet Services Inf. Secur.*, vol. 5, no. 2, pp. 47–64, May 2015.
[38] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins, "#FluxFlow: Visual analysis of anomalous information spreading on social media," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 1773–1782, Dec. 2014.

VISHAL SHARMA received the B.Tech. degree in computer science and engineering from Punjab Technical University in 2012 and the Ph.D. degree in computer science and engineering from Thapar University in 2016. He was with Thapar University as a Lecturer in 2016. He is currently a Post-Doctoral Researcher with the MobiSec Laboratory, Department of Information Security Engineering, Soonchunhyang University, South Korea. His areas of research and interests are 5G networks, UAVs, estimation theory, and artificial intelligence. He is a member of various professional bodies and the past Chair of the ACM Student Chapter-Patiala.

ILSUN YOU (SM'13) received the M.S. and Ph.D. degrees in computer science from Dankook University, Seoul, South Korea, in 1997 and 2002, respectively, and the Ph.D. degree from Kyushu University, Japan, in 2012. From 1997 to 2004, he was with THINmultimedia Inc., Internet Security Company, Ltd., and Hanjo Engineering Company, Ltd., as a Research Engineer. He is currently an Associate Professor with the Department of Information Security Engineering, Soonchunhyang University. He is a fellow of the IET. He has served or is currently serving as a Main Organizer of international conferences and workshops, such as MobiWorld, MIST, SeCIHD, AsiaARES, and so forth. He is the EiC of the *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*. He is in the Editorial Board of *Information Sciences*, the *Journal of Network and Computer Applications*, the *International Journal of Ad Hoc and Ubiquitous Computing*, *Computing and Informatics*, the *Journal of High Speed Networks*, *Intelligent Automation & Soft Computing*, and *Security and Communication Networks*. His main research interests include internet security, authentication, access control, and formal security analysis.

RAVINDER KUMAR received the Ph.D. degree in computer science and engineering from Thapar University in 2015. He is currently an Assistant Professor with the Computer Science and Engineering Department, Thapar University. He has already developed a complete working project on speech recognition and handwritten recognition for Indian regional language (Punjabi). His area of research includes theoretical and practical aspects of combinatorial optimization, approximation algorithm, and mathematical programming. He is the member of various professional bodies and serves as a Reviewer to many referred journals.