International Journal of Security and Its Applications
Vol.9, No.7 (2015), pp.109-118
http://dx.doi.org/10.14257/ijsia.2015.9.7.10

Anomaly Recognition in Online Social Networks

Ashish Rawat1, Gunjan Gugnani2, Minakshi Shastri3 and Pardeep Kumar4

Department of Computer Science and Engineering, Jaypee University of Information Technology, Solan, Himachal Pradesh, India1,2,3,4
{ashishrawat031, gugnani.gunjan2, shastri.minakshi33, pardeepkumarkhokhar4}@gmail.com

Abstract

The popularity of social networking sites has increased throughout the decade and everything that gains immense popularity with great human involvement also brings many challenges and issues along with it. Similarly the excessive use of online social networking causes a great increase in anomalies. In social networking the anomalies are like fake account, account hack, identity theft, spams and many other illegitimate activities. It is thus necessary to detect such anomalous and suspicious behavior of any user at these social platforms, as they could have an adverse impact on users, especially on teenagers. In this paper, we propose various methodologies for early detection of suspicious and anomalous activities. We have done the analysis of various parameters of social networking and its graph like indegree, outdegree, active time of a node (user) and its behavior.

Keywords: *anomaly, anomaly detection, hacking, spam, indegree, outdegree*

1. Introduction

Social networking is providing their users a big and convenient platform to exchange their views, ideologies, share their stuffs and communicate with people around the world. But along with this extreme openness and convenience, social networking is also a very delicate environment that can be easily used to spread unsocial and uncivilized activities which we are considering as anomalies.

In general an anomaly is any kind of irregular or perverted behavior of a node/user from the usual one and anomaly detection in social networks is the identification of such irregular and different activity or deviated behavior which does not conform to a social activity. Anomaly detection aims to find an observation that deviates so much from other observations as to raise suspicion that it was generated by a different mechanism [1].

The size of social networks is increasing rapidly. The bar chart in Figure 1 represents the number of registered users on various social networking websites [2]. It is observed that there are billions of users around the world that are connected to one or more social networking site(s) and this number will definitely grow in the coming time. Such a billion of number covers every age group, whether young or adult. Out of these billions, almost millions of the people are now addicted to these social sites.

The anomalies in social networking are online fraudsters, sexual predators, spammers, hackers, fake accounts, to bully, to plan a terrorist activity etc. The foremost victims of these attacks and threats are the teenagers. Because of their immature and careless attitude, intentionally or unintentionally they fall prey to the attackers and became vulnerable to these kinds of frauds, attacks on social networking. Therefore, it is very crucial to detect these anomalies as early as possible to prevent users from potential disaster and attacks.

The major problems that still need to be addressed in social networking is sharing of vulgar content, promotional fake account and hacking of account.

ISSN: 1738-9976 IJSIA

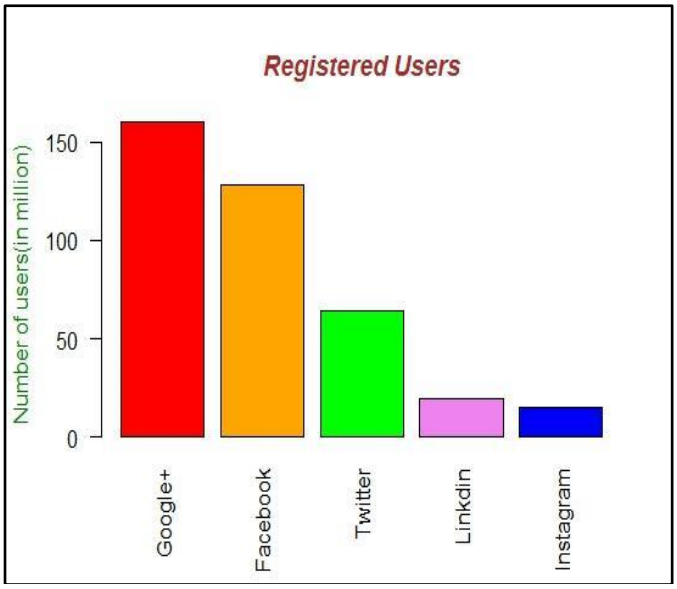

Figure 1. Registered User

People or users share vulgar content on social media like Facebook, Instagram, Google+ etc. which is illegal. More-over, users generally share this kind of inappropriate content without revealing their true identity, which is again a cyber-offense. To abolish these kinds of offensive acts we propose our first methodology which detects such accounts and profiles involved in such activities. By the proposed methodology we will be able to detect suspicious accounts and then the corresponding networking site can take actions against them.

The second problem that our work addresses is hacking accounts. Hacking and changing the password are the conventional hacking approaches which are now being handled efficiently through robust systems. However with time, hackers have become smart they hack the account, extract the useful information without even letting the owner of the account know that his or her account has been hacked. In these types of attacks what attacker generally does is monitor the activities of the account owner and also keeps his eye on user's activities. In this way the attacker will get to know about the personal life of the account owner, his or her close friends and other potential information that may be beneficial to the attacker. In other words information loss could harm the account owner on various grounds and uplift the rival. By doing such an act the attacker not only encroaches upon the privacy of the user but can also use his or her information for some offensive acts. These type of anomalies are generally created by some terrorist organizations, spies, stalkers etc. This kind of attack can be a big threat to the user. So for the protection of the user from these kinds of attacks we proposed another methodology to detect these types of accounts so that the corresponding networking site can take the proper preventive measures.

The third issue that we are dealing here is fake promotional accounts created for promotion of local business, event, website link etc. In addition to this the foremost problem in practice nowadays is promotion of a particular page/timeline on online social media. This kind of promotion includes sharing of post in various groups and timelines, picture tagging etc. Such promotional accounts come under the category of spammers who mock other users. To detect such accounts and to deal with this elicit issue we proposed our third methodology.

This paper is organized in six different sections. In section II, we will be discussing about the previous work done on the anomaly detection in social networking and also includes various techniques and methodologies adopted by the different authors. In

section III we will be describing about the proposed methodology to address the different issues framed above. The section IV is the condensation of experimental setup of our methodology. Ahead in section V, the observations and analysis of the proposed approach is depicted, in addition some graphs representing anomalies are also shown in this section.

2. Related Work

In this section we are going to discuss about the previous or related work done by the authors under the area of anomaly detection in social networking. Most of the papers we have covered are making use of machine learning -supervised and unsupervised techniques for anomaly detection.

J. Silva and R. Willett [3] discussed about the detection of meetings anomalies in social networks. They have proposed a hypergraph based approach which detects anomalous behavior through density estimation on the hypercube.

Salvatore Cantanese, Pasquale De Meo, Emilio Ferrara, et al. [4] has introduced ad hoc Facebook crawler and the Log Analysis tool for exploration and conception of growth and evolution of online social networks. The ad hoc Facebook crawler has been introduced to fulfill the gradually strict terms of the Facebook end-user. Whereas LogAnalysis tool gave a graphical conception of key graph theory and social net-works analysis notions: degree distribution, diameter, centrality metrics, clustering coefficient computation, and eigenvalues distribution.

Nitesh kumar and Ranabothu Nithin Reddy [5] have introduced a framework for automatic detection of fake profiles using classification techniques like Support Vector Machine, Naive Bayes and Decision trees to classify the profiles into two classes, i.e. fake or genuine one.

Q. Cao, M. Sirivianos, et al. [6] have introduced a new tool in OSN operators, named SybilRank . It depends upon social graph properties to rank users according to their probability to be fake. The tool SybilRank proposed by the authors is computationally proficient and is flexible to graphs with hundreds of millions of nodes.

Bimal Viswanath, M. Ahmad Bashir, et al. [7] have discussed about an unsupervised learning technique- PCA (Principal Component Analysis) that models the behavior of normal users accurately and identifies most significant deviations from its anomaly.

3. Proposed Work

The previous work deals with some specific kind of problems in online social networking (OSN) and provided suitable solutions. Still some major issues need to be addressed which we have focused in this paper and these are as follows:

3.1 Inappropriate Content Share

The biggest issue to worry about online social networking is sharing of vulgar, inappropriate, obscene and even pornography content. This content may occur in any kind of form like an image, any video, any message, or website link that is diverting to some inappropriate content. These kinds of contents have a bad impact on young blood and could harm the society and culture. This issue becomes more severe with teenagers as they get influenced early and may deflect to the wrong paths. So it is very crucial to put a hold on this kind of content sharing on widely spread social media.

Users generally share this kind of inappropriate content by using some false identity because users usually do not perform any illegal work by disclosing their true self. So these kinds of users are doing two offensive things one is sharing of inappropriate content and the other one is creation of fake accounts by not revealing their true selves.

So to catch this kind of activities we need to find the loophole since such kind of inappropriate things whenever shared on social media starts getting several

thousands of views in no time. Particularly if we take the example of Facebook, then these kinds of inappropriate content on Facebook have more views than that of likes, comments or shares on that particular inappropriate post. The second loophole is when any account is created with the aim of such activities then their true self is not disclosed which leads to less number of friends. Moreover, people who used to view such kind of content donot add these fake accounts on the list of their friends, users often just visit these inappropriate sharing accounts and view obscene content.

So the key concept we are using here to catch these kinds of accounts is by analyzing their indegree and the number of friends and followers in the friend list of that account. Generally, this type of accounts contains very less number of friends in their friend list and their indegree is very high.

Indegree in a graph network is the number of incidence links on a node. So in a social network graph the number of activities that incidence an edge on a node, is its indegree. The key consideration here is, for a node/user, friends and followers are responsible for the count of indegree on that node/user. Now when we analyze a particular node that has very less number of friends and its indegree is very high or the indegree is increasing very fast as compared to number of friends then we can say that there may be some suspicious activity taking place on that node/user account. Because in general, the indegree of a specific node depends on the network built by a node through the number of friends and followers he/she has. But in case of some defaulter the indegree is significantly very high instead of possible usual indegree as per networks built by linked accounts.

Algorithm 1 Inappropriate content share
Require: OSN graph
Ensure: User Indegree and Number of Friends.

1) Analyze the Online Social Networking Graph.
2) Computation of Outdegree on a node/user.
3) Count of Friends (including followers)
4) Devise relationship among user Indegree and number of Friends.

These activities can be sensed as described above and after detection its the responsibility of corresponding social networking sites to take proper control measures for that particular account and activity. The measures must be like strict identity verification, recognition of account by friends and some restricted activity permitted etc., can be done.

The graph in Figure 2 has been created in the Microsoft NodeXL toolkit. Nodexl: It is an extensible toolkit for large graph networks overview, discovery and exploration implemented as an add-in to the Microsoft Excel 2007 spreadsheet software [8]. It intends to make networks analysis tasks easier to perform. To extract the data from social networking sites we need authorization from that particular site and then we can build the networks and analyze it easily. This graph represents the user/node, at center, and its connections to other nodes while the red colored links representing friends of the user and remaining links coming from others.

The graph is connected graph representing various nodes while edges in the graph representing in/out degrees of nodes. The edges differentiated in red color is to show the edges which are coming from friends and followers to the node at the center. Since edges are coming to the central node from nodes (other than friends and followers) in glut, this could be the sign of anomaly or any suspicious activity at the central node.

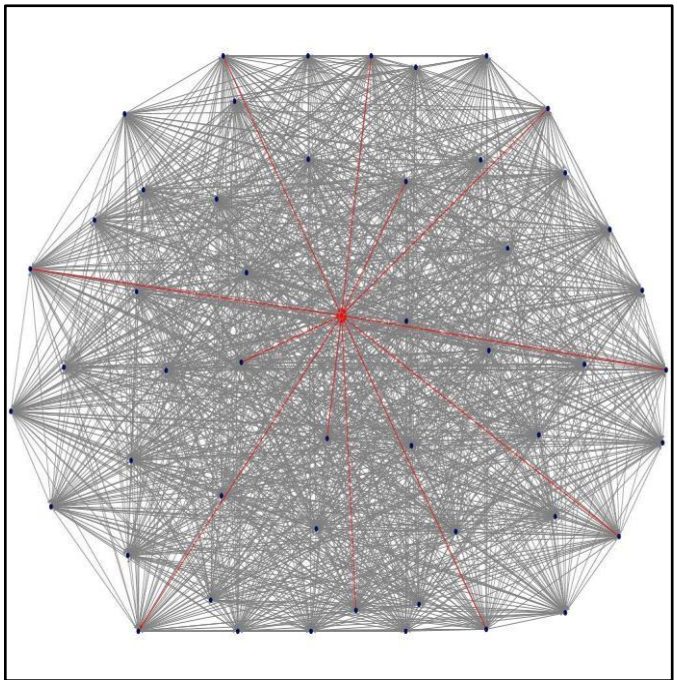

Figure 2. Online Social Networking Connectivity Graph

After computing the indegree and the number of friends and followers, we plot a graph between two and analyze the graph. Now for an account which has very few friends (or nodes attached to it) but its indegree is significantly very high, hints suspicious/anomalous activity.

3.2 Silent Hacking of Account

The second issue which we are trying to resolve here is hacking of the accounts. Nowadays it is very common that hacker hacks the account to just monitor the user activities. This is very common in the cases when the account is hacked by some terrorist or by some spy or by some stalker. In this hacking they do not trouble the user visibly like by changing passwords or by sending spams, etc. In this hacking, hackers invisibly (without letting anyone know about that his or her account has been hacked) just observe and monitor the activities of the account owners on his social networking account. In this manner hacker enters into the private zone of the user by breaking the privacy and monitor activities to get the details of his/her lifestyle and other potential information that a hacker may use in wrong manner. Both the acts hacking and using someones personal information in wrong manner are illegal. So to stop this kind of offensive act we came here with a solution. In these kinds of hacking what pattern we are generally observing is that on hacking, all of a sudden the active duration of a node increases visibly, but the outdegree remains the same as in the past and this type of pattern is observed for a significant period of time. The outdegree is the number of tail endpoints adjacent to a node in its graph.

Algorithm 2 Silent hacking of account
Require: OSN graph
Ensure: Outdegree and Active time relationship.

1. Analyze the Online Social Networking Graph.
2. Computation of Outdegree of a node/user.
3. Devise relationship among user Outdegree and Active time.

For a node active duration is higher than that of before (or in the past), so a sudden increase in active duration here symbolizes an anomaly in the behavior of a node. For example a user that used to remain active on his social networking account for four to five hours a day is now active for around ten or eleven hours a day, but the number of activities performed by the user is the same and we observe the same pattern regularly. Then that account must be there in the list of anomalous accounts and the corresponding site must take proper actions to check this anomaly.

We incorporated a bar plot graph between outdegree and the active time for a node/user in Figure 4. The bar plot represents a relationship between outdegree and active time where the width of a bar represents outdegree while the height represents the active time. Now normally the height and width (for a user having a strong relationship between activities and active time) of the bar are in accordance and it's the pattern of normal user activity. But if the bar plot graph starts showing some highly deviated relationship in height and width of the bars, regularly, then we can come up with the view that there may be some anomaly propelling on that account. The high active duration in comparison to outdegree of a node is responsible for the categorization of an account as anomalous. This methodology works more efficiently for social networking sites which do not involve chatting with their features. The measures that can be taken by corresponding website are- notify the user about suspicious activity seen on his/her account, ask the user to review various locations of access or ask the user to change the password etc.

3.3 Fake Promotional Accounts

The third problem that we are trying to resolve in here is, nowadays fake accounts are created for any promotional events, fake reviews, publicity and other similar purposes. In this, what these users do is create a fake account and try to publicize and promote any celebrity, product, place, etc. Even sometimes the accounts are just created so as to intentionally mock some other users. In this case fake users/nodes do most of their activity on a particular node only and hence we observe that outdegree on a particular node is very high as compared to other nodes which causes anomaly in the behavior of a user. For example, if total outdegree for a node is ten and out of which eight degree is on some specific node and remaining two degree is on some other nodes. This causes the user/node to fall under the category of anomalous activity. In this case we are assuming the outdegree to be the sum of number of likes, posts, shares and tags (where one tag is equivalent to one out degree). The tag parameter is very important here because for promotions and publicizing users tag their friends, or for mocking purposes too.

Algorithm 3 - Fake Promotional account
Require: OSN graph
Ensure: Highest Outdegree on any node.

1) Analyze the Online Social Networking Graph.
2) Computation of Outdegree on a node/user.
3) Computation of highest Outdegree on any node.
4) Comparison among user Outdegree and Highest Outdegree on any node.

We excluded the accounts of page admin, page promoter, page analyst which have authorized roles on the page from our domain of anomaly detection. We again plot a barplot graph shown in Figure 5 which represents the outdegree of the node/user where blue color represents the outdegree on a particular node while red shows outdegree of the rest of the nodes and together both represent the total outdegree. Now an anomalous account activity is the one in which the node is having very high outdegree of a particular node and very low outdegree on some other nodes that says the node is performing too many activities on a particular timeline or page and very less activities on some other timelines/pages suggesting a fake promotional account. Outdegree here, we considered is

by number of likes, share and tags done on a page or timeline. The measures that must be taken by the respective website like strict identity verification, recognition of account by friends and some restricted activity permitted etc.

4. Experimental Setup

The proposed work has been implemented in R, R is a programming language and software environment for statistical computing and graphics. The source code for the R software environment is written primarily in C, Fortran, and R. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. We have used 64 bit operating system windows 7 having processor with 2.30GHZ frequency working on 4GB RAM. NodeXL tool has been used in this work to explore the graph network. NodeXL can work on Windows 7,Vista or XP and also have proficient working with Office 2007,Office 2010 but still not tested with Office 2013.

We need dataset with a combination of original and fake profiles accordingly. Since there are no such datasets available because of privacy issues in various social networking sites or online social networks. So we need to prepare the datasets by randomly walking the profiles from social networking sites. We have used the 500 profiles for our datasets.

5. Observation and Analysis

We have performed computation and analysis which involves iterative strategies on different kinds of data sets. Different nodes/users have different activities on social networking sites, so we have to deal with variety of datasets. Based on behavior of data sets and the elicit problems in the introduction section, we opt different strategies to detect the anomalous and suspicious behavior.

5.1 Sharing of Inappropriate Content

For the anomaly of fewer friends and significantly high indegree which could be a result of inappropriate content share, we plot a graph between two. Note that we are not aware of strategies and algorithms used by any social networking sites as they are hidden for security reason.

In our approach we compared the number of friends and the indegree at a node, since the indegree at any node depends upon the span of users networks, which further depends upon its friends and followers and their friends. We have compared this relationship for many nodes, which found to be working usual at several nodes, but the behavior of few nodes foundvery unprecedented on this context. As depicted in Figure 3, the indegree of the node in unprecedentedly high with very few friends on followers linked to the profile.

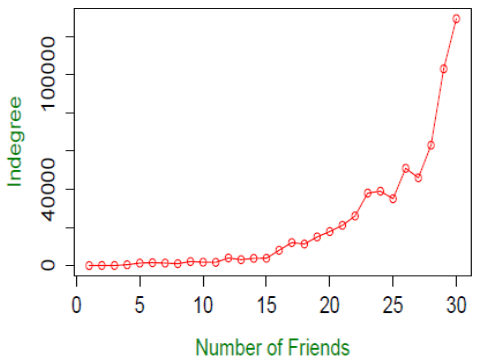

Figure 3. Number of Friends vs. Indegee

5.2 Silent Hacking of Account

For the silent hacking of online social account, we have compared the active duration of a node to the outdegree of that node. After processing the sample datasets we have found some of the anomalous accounts in which there was huge variation in the relationship active duration of past and present, but along that outdegree was almost equivalent to the past and present. Hence some accounts are lying under the anomalous category.

In the barplot shown in, Figure 4, the width represents the outdegree and height represents the active time, both have an intuitive relationship in normal days but the bars representation of last week shows significant variation than past days. This variation reflects anomaly in the behavior of the node. This anomaly could be the result of account hacked and monitored by some malicious entity.

5.3 Promotional Fake Accounts

Promotional fake accounts are built for biased promotions, fake reviews that misguide and mislead the other users. So to detect these breeds of accounts we applied the above mentioned approach to the sampled data and found that some of the accounts are suspicious. They fall in the suspicious category because some nodes have significantly high outdegree on a particular node only.

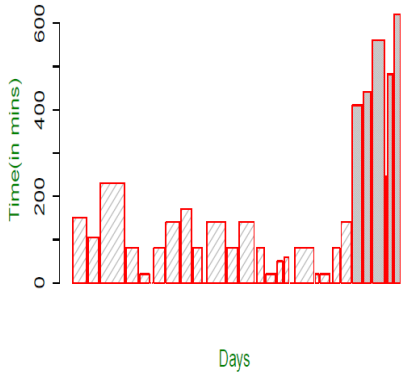

Figure 4. Outdegree vs. Active Time

The barplot in Figure 5 is representing the day-wise outdegree of a node/user say node-xyz where blue color is representing the outdegree on a particular node (with highest indegree from node xyz) and red color shows the outdegree to the rest if the nodes while together both shows the total outdegree of node-xyz. Now it is can be clearly visualized in the bar graph that outdegree on a node is significantly higher incomparison to total outdegree and which is the case of anomaly and could be a fake promotional account.

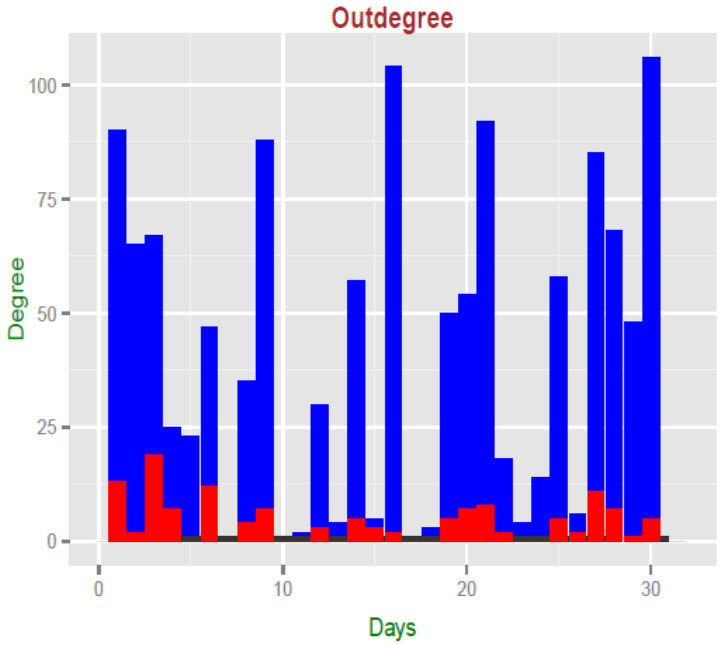

Figure 5. Outdegree

6. Conclusion

With the immense use of social online sites, we face many significant challenges like continuous creation of fake user accounts, hacking of accounts and other illegitimate acts which fall under the category of anomalous and suspicious behavior. So we propose three methodologies to cope up with these issues. First methodology involves the comparison between the number of friends and indegree for detection of anomalous behavior of a node e.g. sharing inappropriate content. The second methodology composed of comparison of outdegree with active duration of a node for detection of account hacking. Now for the detection of fake promotional accounts we have third methodology which is comparing the highest outdegree on any node, by the user under examination, to the total outdegree of the user. These three proposed approaches try to figure out the unusual behavior/anomaly of a node/user.

References

[1] D. M. Hawkins, "Identification of outliers", Chapman and Hall **(1980).**
[2] Z. Whittaker, "Social Media 2014 Statistics". Digital Insights", **(2014)** June, Retrieved June 2014, http://blog.digitalinsights.in/socisl-media-users-2014-stats-numbers/05205287.html
[3] J. Silva and R. Willett, "Detection of anomalous meetings in a social network", In 42nd Annual Conference on Iformation Sciences and Systems, CISS March 19-21**(2008),** pages 636641, (**2008**).
[4] S. Cantanese, P. De Meo, E. Fer-rara, G. Fiumara, and A. Provetti, "Extraction and analysis of facebook friendship relations", Technical report, University of Messina, Italy and University of Oxford, UK, **2010.**
[5] N. Kumar and R. N. Reddy, "Automatic Detection of Fake Profiles in Online Social Networks", Diss. (**2012**).
[6] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services", In Proc. of NSDI, (**2012).**
[7] B. Viswanath, M. Ahmad Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishna-murthy, and A. Mislove, "Towards Detecting Anomalous User Behavior in Online Social Networks", In USENIX Security, (**2014**).
[8] M. Smith, B. Shneiderman, N. Milic-Frayling, E. M. Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer and E. Gleave, "Analyzing (Social Media) Networks with Nodexl", Proceedings of the 4th International Conference on Communities and Technologies. Springer, Berlin **(2009)**.

Authors

Ashish Rawa, he has received his B. Tech. degree in Information Technology, in 2012, recently he has completed his M.Tech. From Jaypee University of Information Technology and his area of interest includes Web Mining, Anomaly Detection in OSN.

Minakshi Shastri, she has received her B. Tech. degree in Computer Science and Engineering in 2012 from Himachal Pradesh University Shimla (H.P.), recently She has completed her M. Tech. From Jaypee University of Information Technology and her area of interest includes Data Mining, Anomaly Detection, Bayesian networks.

Gunjan Gugnani, she has received her B. Tech. degree in Information Technology, in 2013 from Jaypee University of Information Technology, recently She has completed her M. Tech. From the same university and her area of interest includes Cloud Computing and Cryptography algorithms.

Pardeep Kumar, he obtained his B. Tech. (Information Technology, June 2004) from Kurukshetra University, Kurukshetra, Haryana. He obtained his M. Tech. (Computer Science & Engineering, May 2007) from Guru Jambheshwar University of Science & Technology, Hisaar, Haryana. He has completed his Ph.D. (Computer Science and Engineering, Nov. 2012) from Uttarakhand Technical University, Dehradun. His interested research areas are Computational and Machine Intelligence, Machine Learning, Evolutionary Computing, Information Retrieval.