Community Spam Detection Methodologies for recommending nodes

J.Jeyasudha [1], Dr.G.Usha[2]

Abstract*: The most popular and leading social network service online now days is Facebook, twitter and Linked In. When socializing becomes usual, the probability of threats and unwanted posts (Spams) comes naturally. To identify and block such Spams, there are a few techniques available recently. However, the efficiency of such tools to combat with spammers seem tedious due to the public unavailability of critical pieces of Facebook Information like Profile, Network Information, Posts and more. Literature shows that there are many researches been carried out to find and combat malicious accounts and spammers over last two decades. In this paper, a review of similar methods that works with detection of spammers in a community on Social Networking Website with the help of mindmap that is given. The work is comprehended in how data is collected, types of spammers, classifiers, machine learning, review on spammers and community detection and whether it is graph based or non graph based dataset. A survey of research publications on Spammers and Malicious account based on malicious categories for the detected communities with the help of various categories discussed in the mindmap.

Index Terms: Social Spam, Community Detection, Influential Node.

I. INTRODUCTION

Social media thus is an evolution of the Internet, where people connecting themselves with the world. The most important types of social media span are, Bookmarking sites, Blogs, RSS Feeds, Linking and posting, Micro blogs Content Rating, Widgets , Audio podcasting and Video podcasting, Social Networking. A social network web site allows a user to get an user account to create a digital authority of themselves ,secondly to choose members of the social media to get connected and engage with these users, then use an interface (API) to build applications "the information a social network collects about a user" includes contacts, where they are located, associations, personal information, their history of work, personal preferences, who you're friends with, etc

The 82 percent of the majority people in the world engages in social media weekly once , with half of the people participating every day(48% users). One in six (16%) use social media to get information about an emergency. In the Figure 1 represents how many users are using the social networks are illustrated, facebook as whole is having many users. During an emergency, nearly one third of the people population would use social media to let others know they are safe. Face book is a podium to share news, requests for feedback, queries, and links with an engrossed community that help people a place to share information with each other. Face book contains People-based, groups, or webpage-based accounts and average user spends almost 3 hours per day on Facebook.

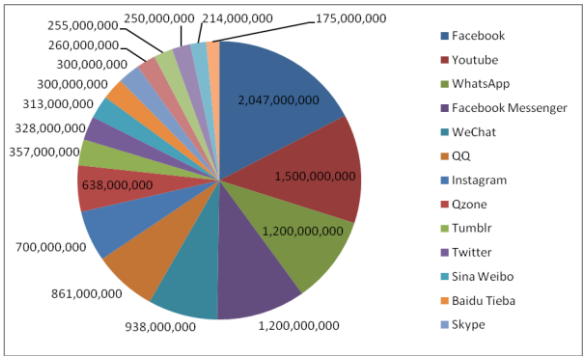

Figure.1 User Accounts in Different Social Media Networks

In section 2 it explains about the online social network and the section also explains about the mindmap of the community detection and the review of the various community detection. Once the communities are detected the spammers inside the community van be found out easily so the section 2 also explains about the various spam detection. and section 3 concludes about the summary of the various spam detection and the community detection.

II. ONLINE SOCIAL NETWORK (OSN)

Online social networks (OSNs) (Figure 2) have developed as vital platforms for people to commune across the world. After introduction of very first Social Network SixDegrees in 1997, several social networking platforms such as Facebook, Twitter and LinkedIn have been developed and became popular [1]. Advancement in Mobile Phones and Computers pushes the social network to strive for new developed applications for socializing and for fun. Moreover, corporates use online applications and features to brand and market their products which in turn results in more number of online user registration every day. As an outcome, an individual has a minimum of 10 to 15 online user accounts to make a living now days [4]. On the other hand, Celebrities are also using online social media to communicate with their fans. Whereas media newsprints also started using online social media as their play ground to promote and distribute their content and services. This makes a scenario that, an individual's data is present across the globe with or without their knowledge. That creates a platform for malicious user accounts and spammers.

Social media sites have both sensitive and insensitive datas, friends lists,family,and contacts; logs of actions, priorities, and favourites, location maps to find areas and how regularly; time stamped posts that point to where a person

Revised Manuscript Received on July 05, 2019.

J.Jeyasudha, Department of Software Engineering, SRM Institute of Science & Technology,Kancheepuram,India.

Dr.G.Usha, Department of Software Engineering, SRM Institute of Science & Technology,Kancheepuram,India.

was and when; and the content of the posts themselves, where people detail their thoughts, feelings, and ideas. Spammers use social engineering attack, malware and spam to steal credentials of legitimate users and compromise their user accounts so that they can deceive their friends and to spread customized spam messages [5]. At this moment, the privacy of online user and maintain the same has become a major concern in online social networking.To reduce these activities online social networks poses a method to differentiate human efforts from other automated activities. For that, CAPTCHA has been introduced. However, this idea has a limitation over identification of clone attack and allowing spammers to gain access over legitimate users data and posts. The next method used for combating spammers was blacklisting, which verified against URL posted by a user with popular APIs such as Google Safe Browsing and PhisTank. Since, the time taken for comparing an URL against APIs data set is too large, approximately 85 % of the visitor accessed the spam URL before it is avoided.

Academic and Industrial researchers have proposed alternative methods. To identify the threat Facebook proposed immune system, EdgeRank algorithm provides a score to each user based on their fair usage of features [10]. This has a limitation of spammers can plan their activities on Facebook network and boost their EdgeRank score. Whereas Twitter developed a rule of thumb for securing their network and yet again could not stop spams and malicious user accounts. Crowdsourcing method is introduced by Wang et al, which detects the human efforts and identifies fake user accounts on social networks. This approach is best suited for smaller data and not that much successful when data becomes huge, since this requires a lot human effort to get higher accuracy in testing. At this moment, Graph based analysis and machine learning analysis methods were brought in to provide better detections. A friendship invitation graph developed by Yang combines different features that trains machine learning process to differentiate spammers from users. Whereas, the method proposed by Vishwanath et al, that revealed a limit to use only the structure of the social network to identify spammers leads a better machine learning understanding.

ONLINE SOCIAL NETWORK DATASET

The online network dataset is categorized in to two main domains Graph based and Non-Graph in figure 3 based by comparing the previous studies dealt in line with malicious accounts. The graph method uses nodes and edges to model a social network as graph. The non-graph method uses a detection system which is formulated by different features that are extracted from social network data.

Figure 2 Social Media Network

Using Barabsi-Albert preferential attachment model, few researchers developed web crawlers that helps to get the private graph data from social network of importance. These are classified as synthetic social graphs and they assume social media network as scale free model and they follow a power law distribution. This method has a limitation of enabling a password for public non-graph dataset due to the fright of violating users' privacy.

Further, these model have only few number and limited attribute of registered users which in turn difficult for the researchers to develop the model further. This constraint the researchers to use APIs to collect private data by the social network provider using web crawlers.

Manual collection is the best solution for programming issues, but need more manual labor. Data can also be good/bad collected by humans using thoughts rather than computers that cannot detect the target of some subtle human phrasing. Facebook app that does the data gathering for you. The Facebook API ,Twitter Streaming API.Depending on the data you want to get you can connect to the Graph API for example JavaScript, PHP or (my recommendation) R. Crawler,Web crawler embedded in a Chrome extension.Java API "HTML Parser", MyPageKeeper, Honeypot.

An Application Programming Interface (API) is a set of procedures, tools and protocols, it is used for construct various applications and software. Social network platforms offer APIs to users to develop various new web applications. That will benefit its programming structure for outside groups to utilize and create new features to their websites . An API usually consist of an operating system, a web-based system, or a database tool, and always based on a specific programming language.

It is useful for developing applications for the different system. APIs can work as the GUI components, or to access computer hardware or database like the hard disk driver. Through various APIs, third parties and researchers have access to the instant data, user activities, celebrities' actions and the most popular topics in the world. In this section, we will introduce the background information about Facebook API and Twitter API, and the datasets collected during the research and then classify research goal before we analyze the datasets.

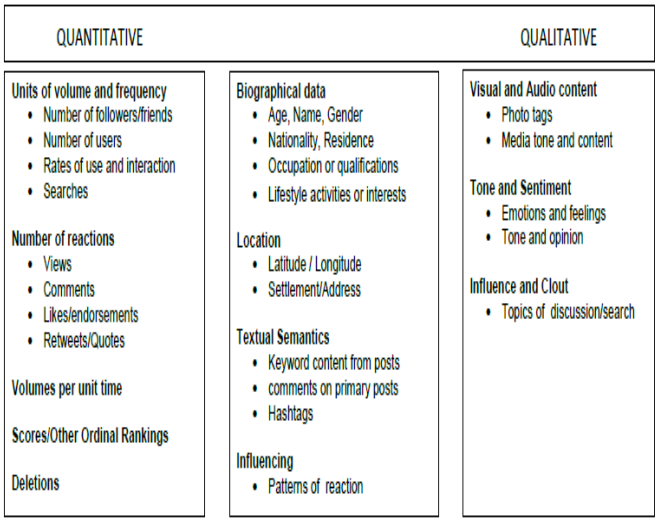

Figure 3 Social Media Analysis

MALICIOUS USER ACCOUNTS ON SOCIAL NETWORKS

There are two categories of malicious activities used in social networks namely "Fraudulent / Career Spamming" and "Compromised User accounts". The sensitive / precious information of victim is obtained from the victim through embedding a malicious link to phishing webpage. With that information of user account owner, his / her friends and friends of friends' database, a fake user account (Sybil) can be created by any spammer and it can be used to spread malicious contents. These fake contents are used to overshadow rightful users and demoralize their belief and relationship in social network so that the spammer can perform malicious activities through legitimate user profile and shown in the Figure 4. These activities include social spamming, private data harvesting and drive by download [12].

Attackers may equip them with automated characteristics which mimics real users to make them look alike real user so that the fraudulent activities can be stretched to a higher time period. Having a fake user account online and making millions dollars has become a prime business now days. Recently it is discovered that there are more fake user accounts in the name of celebrities, politicians and popular organizations [13]. These scenario puts social networks in to lot of risks and strive hard for a solution for the same.

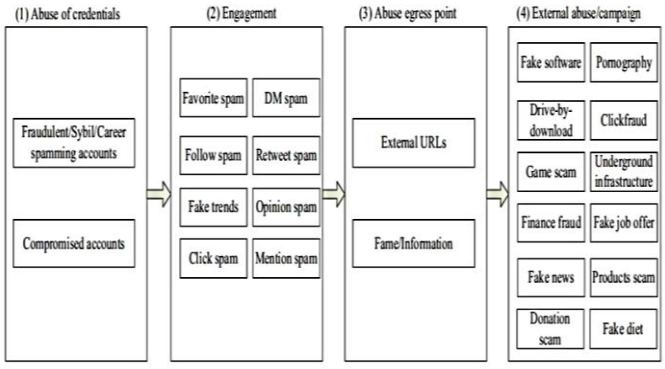

Figure 4 Abuse of Social Networks

The second threat available online is Compromised user accounts which is a hijacked user account of a legitimate user through posting an URL which forces the user to click on it and diverts the page to phishing webpages. Literature shows that user accounts which can compromise are more useful to spammers than spam user accounts which carry the spam. Since compromised user account has more trust and relationship with other users of legitimate user, the chance of leveraging true relationship is higher when compared. And after hijacking legitimate user user account, spammer will start posting malicious contents in legitimate user page. But the study shows that, the spammer could not match the pattern of posting as legitimate users. This creates a scenario of sudden changes in real users posting behaviour. As an example, the victim may be engaged in posting malicious contents involving pornography, donation and sharing related posts. Once these parameter are figured out by combating services, the spammers devise new strategies to overcome the detection approach and make this as a cat and mouse fight.

IMPACT OF MALICIOUS ACTIVITIES IN OSNS

Since malicious user accounts on social network has been increased drastically, the impact of malicious activities are also gone higher. With reference to the report shared by Nexgate in 2013, the amount of spam distribution has risen up to 35 % in the first half of the year. And the report discusses few parameters as follows:

1. At least 5 % of all applications of the social structure are for spam purpose.
2. Malicious user accounts posts large volume and faster content in social network than real user accounts.
3. A spammer distributes malicious content on at least 23 social networks.
4. There are five spammers born for every seven social media user account.
5. 15 % of all social spam message contains an URL that spreads spam.

Literature shows that the number of identity fraudulent cases has reached 13 million per year over the past six years and social spammers cause a loss of $200 million per year to social trust, productivity and profit. As the rise in malicious activities online, it is mandatory to remove fake user accounts that poses threat to legitimate user on the network.

III MINDMAP FOR COMMUNITY DETECTION

The communities or groups in social media, where people are social,

- user-friendly social network help humans to widen their societal in unique ways
- intricate to communicate with friends in the substantial world, and is easier to locate friend in social network with related interests
- communications linking nodes can help determine communities

The MindMap(Figure 5) is done under various categories such as

- Factorizations (nonnegative matrix factorization (NMF) has been widely adopted for community detection due to its great interpretability and its natural fitness

for capturing the community membership of nodes),

- Deep learning (Deep learning also known as deep structured learning or hierarchical learning is

knowledge of the data structures, for the work-specific

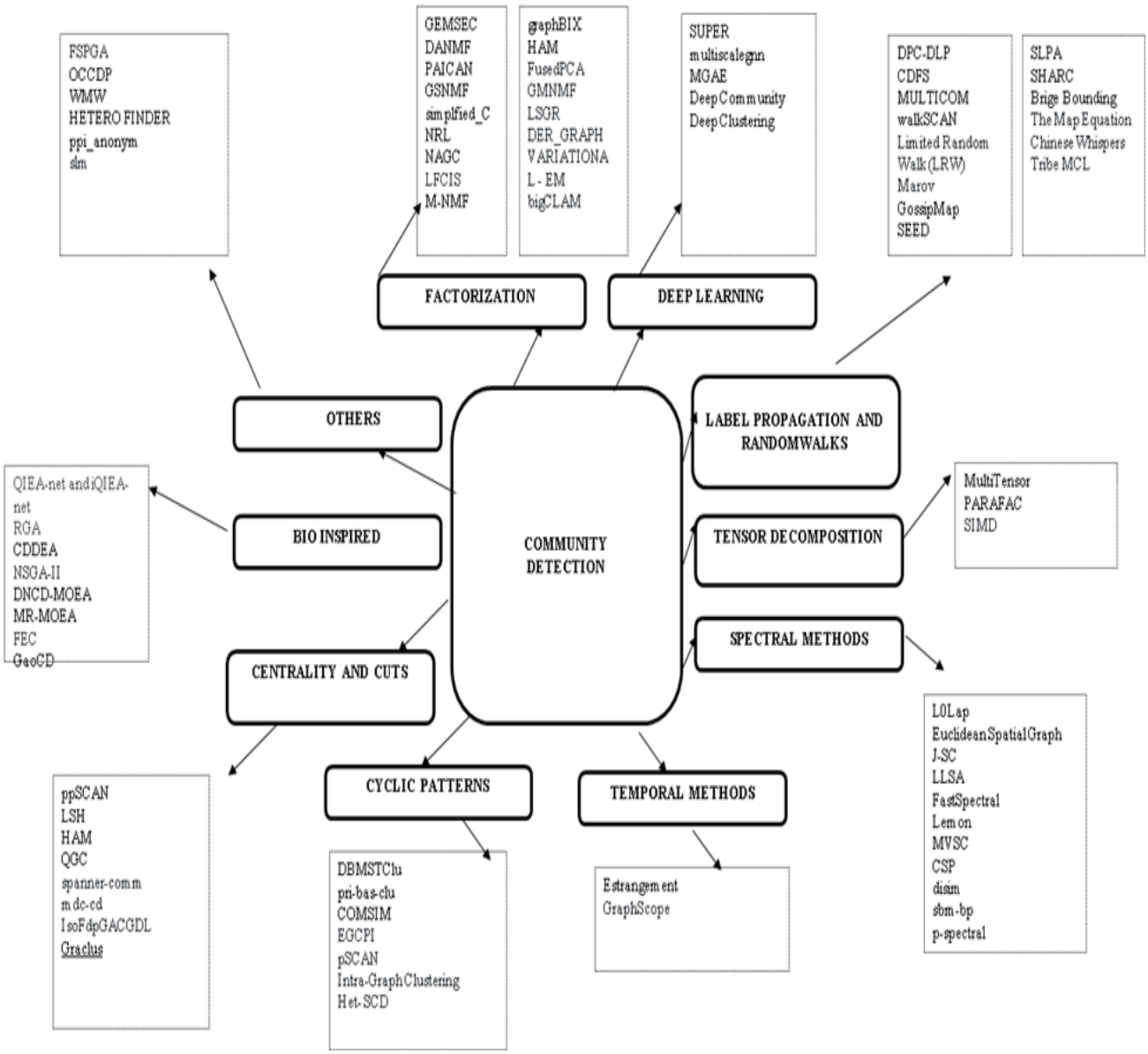

Figure 5 MindMap for Community Detection

algorithms. Learning can be supervised, semi-supervised or unsupervised),

- Label propagation and Random walks (The Label Propagation algorithm (LPA) is a rapid algorithm for finding communities in a graph based structure. It detects these communities with help of the network structure, and no need of any prior information about the communities.)
- Tensor Decomposition, (Tensors are elevated dimensional generalization of matrices. In recent years tensor decompositions were used to design learning algorithms for estimating parameters of latent variable models like Hidden Markov Model, Mixture of Gaussians and Latent Dirichlet Allocation)
- Spectral and Temporal Methods,
- Cyclic patterns, centrality and cuts.
- And some of the methods are categorized under the

bio inspired and physics.

The mind map provides us the overall methods that are involved in finding out the communities. After the Communities are found they are helpful for finding the influential persons easily. Influential persons in a community is easy to find and based upon the metries the communities operate. The communities are evaluated with the help of the many parameters like modularity and NMI,ARI.

IV COMMUNITY DETECTION IN SOCIAL MEDIA

A community is a collection of nodes between where the communications are (relatively) recurrent or discovering groups in a network where individuals' group_memberships are not explicitly given a.k.a. clustering, grouping, finding organized subgroups. If a social media network is given as input, the output will be community attachment of

(some) actors. And it is used in understanding the interactions between people, visualize and navigating vast networks and form the basis for other tasks such as data mining

A review for Community Detection

Methology Name	**Author's**	Methodology	Dataset	Evaluation	Demerits
2004, Fast Modularity [20]	Aaron Clauset	1.Greedily optimizing the modularity. 2.Adjacency Matrix,hierarchical clustering	Amazon.com	Modularity	Use power-law form P(s) ~ s− for some constant
2005, Walker Trap, Communities In Large Networks,[21]	Pascal Pons	1.Similarity measure between vertices based on random walks 2.Adjacency Matrix, hierarchical clustering	Random generated graphs	Modularity	large amount of memory is needed
2005, Spectral Algorithm, [22]	Luca Donetti	1.Eigenvectors of the Laplacian matrix 2.Dendogram hierarchical clustering	Zachary karate club	Modularity	Generalization to the case of weighted networks.
2007, Label Propogation, [23]	Usha Nandini Raghavan	It needs pre-defined objective function nor information about the communities before the processing itself.	1.Zachary karate club 2.US football network	Modularity	Prior information is not available for real world social networks
2007, Modularity Opt(simulated Annealing),[24]	Marta Sales-Pardo	1.an ensemble of organized nested random graphs, Affinity measures and clustering methods	Metabolic networks	Mutual information	Huge pathways are composed of smaller units.
2008, Louvain [25]	Vincent D Blondel	1.heuristic method that is based on modularity optimization	Belgian mobile phone network of 2.6 million customers	Modualrity	Storage of the network in main memory takes more computation time.
2008, Infomap, [26]	Martin Rosvall	1.Directed and weighted graphs are used 2Random walks probability on a network and Huffman Coding is needed	Scientific disciplines: Science, Nature Journals and Proceedings of the National Academy of Sciences	modularity or or cluster-based compression	Direct connections are not available because fields on opposite sides of the ring are associated only through intermediate fields
2009, Potts model, [27]	Peter Ronhovde	1.It manipulates within the replicas of same group for the over a range of resolutions 2.avoids the “resolution limit” 3.weighted Hamiltonian as absolute Potts model	Erdos-Renyi random graph, UCINet for network data	NMI, Modularity and the RB Potts model (RBPM)	Large number of individual community solutions are needed
2009, Propinquity dynamics,[28]	Jianyong Wang	Propinquity is a measure of the probability for a pair of nodes involved in a consistent community structure	Wikipedia linkage graph dataset	propinquity with existing algorithms	It confirms the conditional convergency of propinquity dynamics

2010, Link-Plus,[29]	Yong-Yeol Ahn	1.Reveal overlap within the communities	Amazon.com and PPI networks	Overlap quality, threshold, t	use fine metadata, the quality will remain high
2010, MOSES, [30]	Aaron McDaid	Detects highly overlapping community structure, (with variance in the number of communities each node).	friendship links between students of ve US universities.	Modularity Maximization	Normal overlapping structures are ruled out.
2010, Greedy Clique Expansion, [31]	Conrad Lee	It identifies dissimilar cliques as seeds and uses it in optimizing a local fitness function.	Facebook friendship data	NMI	Multiple scales network is not considered
2010, COPRA(Label Propogation), [32]	Steve Gregory	Label propagation technique lengthen the label and proliferation step to include data about more than one community	Autism bibliographic dataset	NMI	algorithm is highly amenable to parallel implementation
2010, Top Leader [33]	Reihaneh Rabbany Khorasgani	Followers for a influential leader	Karate,football,strike	purity and Adjusted Rand Index,modularity	Number of desired communities are needed.
2010 Skeleton Clustering,[34]	D.Bortner	Requires minimum similarity parameter for the good cluster (not agglomerative)	Enron email dataset	existing algorithms	No automation.simarity parameter needed
2011, (State-of-Art) OSOLOM,	Andrea Lancichinetti	Detect clusters in networks with edge directions, edge weights, overlapping communities, hierarchies and community dynamics	LiveJournal and UK Web Dynamic datasets: the US air transportation network.	LFR benchmark	To reveal the connection between the structures of the system at different time stamps
2011, Multi-Level-Infomap	Martin Rosvall],	Description of the random walker in multilevel. Optimal number of levels for the dynamics on the network	journal citation network of science, the human disease network and the global air traffic network	multilevel benchmark test	The algorithm can only extract the fine-level modules
2012, Consensus Clustering,	Andrea Lancichinetti]	Stochastic methods partition parameters	APS citation network	NMI	Excludes the cluster vertices for computing Jaccard index
2012, Community Affiliation Graph Model,	Jaewon Yang	It builds on bipartite node community affiliation networks.	http://snap.stanford.edu	similarity of the members	Finding that community overlaps are denser than communities themselves nicely extend the notion of homophily
2012 Maximal k-Mutual-Friends,	F.Zhao]	efficient approach to discover cohesive sub-graphs and summarizing textual interactions between social actors as tag cloud	Epinions Twitter DBLP Flickr 1, FriendFeed Facebook DBLP	with existing algorithm	To maintain the cohesive sub graphs with frequently updates
2012 -Matrix Blocking Dense Subgraph Extract,	J.Chen	Matrix column similarities is done by exploiting the links and no need	bloggers of different political orientations,	Clauset, Newman, and Moore (CNM) Approach	Not appropriate for evaluating the partial graphs

		number of clusters			
2013, (State-of-Art) Large Scale CAG, BigClam,	Jaewon Yang	detect densely overlapping, no overlapping communities in massive networks 2.maximizing the likelihood	LiveJournal Friendster Orkut Youtube DBLP Amazon	Running time on the networks is measured for .Non-negative matrix factorization	No sharing between common communities.
2013, Ensemble	Michael Ovelg¨onne	To spot high quality partitions from an ensemble of partitions with lower quality	uk-2002 and uk-2007-05.	3.3 billion edges with 50 node clusters	more overlaps of the ensemble result in high quality core groups.
2013, Fast Algorithm for Modularity-based,	Hiroaki Shiokawa	To find clusters with high modularity graphs of unprecedented size to be processed in practical time	dblp-2010 l,journal-2008 uk-2005, webbase-2001 uk-2007-05	Synthetic graphs by DIGG and BGLL	incremental aggregation contributes most to the improvement.
2014, Combo Optimization,	Stanislav Sobolevsky	Capable of handling various objective functions,	Orange and British Telecom	Modularity optimization with several methods	30 000 nodes on modern workstations with huge network is needed
2014, SCD,	Arnau Prat-Pérez]	Unprecedented size of graphs are processed in short execution times.	benchmark datasets provided by SNAP	NMI and F1Score	SCD is not able to detect overlap
2014, RelaxMap	Seung-Hee Bae	RelaxMap. This algorithm relaxes concurrency assumptions to avoid lock overhead,	directNet-1k directNet-5k directNet-10k soc-LiveJournal1 soc-Pokec wiki-Talk	benchmarking community detection algorithms.	Consistency relaxation feature
2015, GossipMap	Seung-Hee Bae	Formulation of the map equation by rewriting it with sequence of vertex moves, and evaluated incrementally and locally.	Twitter follower network and .uk domain	NMI	Require using multiple machines
2018 Symmetric NMF with PU Learning	Seiji Maekawa	1) it learns a non- linear projection function between the different cluster assignments of the topology and the attributes of graphs 2.leverages the positive unlabeled learning	WebKB Citeceer cora	Symmetric NMF	Adjacency and attribute matrix
2018PAICAN	Bojchevski and Gunnemann	1.This method achieves high clustering quality after removing anomalies it detects 2.performs ¨ anomaly detection and clustering on the attributed graph at the same time	Lawyers,Parliament,Social Papers,cora	NMI	it can only handle categorical attributes.

201 7JWNMF	Huang et al	factorizes both the topology and the attribute matrices at the same time	Real Time datasets	NMF	JWNMF uses two model parameters λ and A for adjusting attributes weigh,the cost of learning A is expensive
2017 Graph Convolution Networks	Kipf and Welling	semi-supervised learning method for a graph, has obtained considerable attention from machine learning and data mining fields due to its high performance in classifying graph vertices	Real Time datasets	NMF	This approach needs a subset of true cluster labels on vertices, and thus its goal is different from that of the attributed graph clustering

Table.1, Review on Community Detection

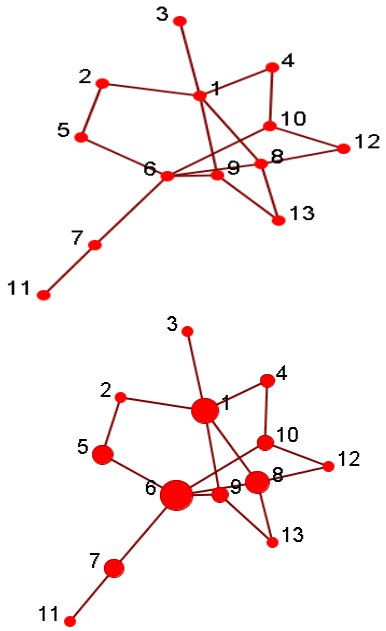

Figure 6. Centrality Analysis/Influence Study (Top 5 influential nodes 6, 1, 8, 5, 10)

The table 1 consists of the various community detection methods based upon the various categories , the fast modurality,walker trap and spectral algorithm[20-22] are of the hierarchical clustering which has a disadvantage of the amount of memory and the constant usage. The methods [23-24] is used based upon the on information of the users given.The Louvain method is most popular method for the modularity optimization.

The methods [26-30] is based upon the distance of the nodes connectivity and the measures differs for each and every method.The other methods are just the expansion of the methods that are already discussed. The methods PAICAN,Symmetric NMF which talks about the attributes that are used, and JWNMF is based upon the model parameters.The Figure 6 represents the influence study of the nodes. When a community of nodes given from that popular nodes / Influential nodes are found out for the ease of malicious user nodes.

V SPAMMERS ON SOCIAL NETWORKS

A social network structure made of nodes that are connected with other nodes by various dependencies like friendship, kinship, etc. The representation are nodes(members) and Edges(relationships). Various forms of social network structure is Social bookmarking, Friendship based networks (face book, twitter),Blogosphere, Media Sharing ,Folksonomies.

There are many ways to analyze Networks ,they are to Predict a type of a given node by Node classification, to Predict whether two nodes are linked by Link prediction, to Identify densely linked clusters of nodes using Community detection and How similar are two nodes/networks by Network similarity.

This work mainly focusses on spam user account, fake user account, compromized user accountand phishing detection. For that, the variation of each and every category of malicious user accounts has been studied carefully and each category of malicious user account has been grouped. From online repositories like IEEE, ACM, ScienceDirect and Springer, the article search were performed and the results discussed herewith.

Literature shows that there are many algorithms developed to identify malicious user account and only few of them discuss the past developments made in the area of malicious user accounts detection and spammers control.

The review of the spammers in the table 2 is based of the datasets, metrics, data extraction method, classifiers, account type and the dataset.The dataset that are considered for the review is mainly extracted using API,crawler or any random code from two social media Facebook andTwitter.

Now a days collecting the facebook data was little bit diffciut due to world user account issues and the token given for collecting the data is one for per day. And the Twitter data's can be easily downloaded using any API. Mostly the spam detection is done for the twitter dataset.

The spammers are categoried based upon the fake profile,inactive accounts and the URL based spammers.Some of the spammers attach the content in the photos /Videos they share within the closed group.The features that are used for the Twitter are mostly of the Text features and social features,

- Followers count ,
- People Following,
- Account age,
- FF Ratio,
- Total Tweets,
- Hash tag ,
- Frequency of Tweet,
- in/out degree,Betweeness

- counts in a message,
- comment, post was shared/not ,
- tagged people count ,posted time

The features used in the facebook/Twitter are On-demand features,Aggregation-based features, Generic statistical features, User-based and Content- based features, Text based features. The classifiers used for the Training and Testing the data will J48, Decorate and Naive-Bayes, Random Forest, bootstrap aggregating, or bagging, K nearest neighbors, Bayesian, Support Vector Machines, SVM, KNN,Logistic regression , Latent Dirichlet Allocation, Decision Tree. The metrics that are used to Review of Methodologies for Spam Detection

evaluate the trainign and testing samples are Accuracy, MCC, F-Score, Sensitivity and AUC.,the below review shows the accuracy ,F! metric score between the range of 90 to 100.

The review helps us to find out what are the features used and correct classifiers for features.The count of the data if it varies high the classification has to be done with the help of hadoop and some of the latest techniques of the deep learning.

The factorization method of community detection will help the huge amount of data with the help neural networks.

3

Method	Name of the Social Media	Spammer Type	Data Extraction Method	Features used	Classifiers	Accuracy, MCC, F-Score, Sensitivity and AUC.	Labelled Data Count	Year
A hybrid approach for spam detection for Twitter[1]	TWITTER	Spam user account	API, Guofei Gu	Followers count , People Following, Account age, FF Ratio, Total Tweets, Hash tag ,url,Metion's ratio, Frequency of Tweet, in/out degree,Betweeness	J48, Decorate and Naive-Bayes, Random Forest	92 % Accuracy	400k tweets	2017
A Machine Learning Approach for Twitter Spammers Detection[2]	TWITTER	Fake user account	Twitter Streaming API,Crawler	URL count on each tweet ,portion of URL tweets,portion of replied tweets, portion of spam tweet words ,Time between posts (mean)	Random Forest algorithm, bootstrap aggregating , or bagging	Fmeasure-92%	54981152 user accounts	2014
Spam Detection for Closed Facebook Groups [3]	FACEBOOK	To Find spam in closed groups	The crawler in a Chrome extension.	Text features and social features, words count , spam words features count, URL count in a message, embedded videos count, any message contained an image, likes count for a message, hashtags counts in a message, comment, post was shared/not , tagged people count ,posted time	Random Forest machine learning algorithm	98% -EFFICIENCY	1200 LABELLED POSTS	2017
Detecting Malicious Facebook Applications[4]	FACEBOOK	Distinguish malicious apps from benign ones.	Facebook API	On-demand features,Aggregation-based features	FRAppE Lite's classifier.	99.5% accuracy,	2.2 million users on Facebook.	2012
A Generic Statistical Approach for Spam Detection in Online Social Networks[5]	TWITTER AND FACEBOOK	SPAM PROFILE	Java API "HTML Parser",	Generic statistical features to identify spam profiles.wall posts, fan pages and tags,tweets, mentions or hashtags	Jrip, Naıve Bayes, and J48,	95 % FOR FACEBOOK AND 97% FOR TWITTER	320 Facebook profiles,	2013

Community Spam Detection Methodologies for recommending nodes

Intelligent Twitter Spam Detection: A Hybrid Approach [6]	TWITTER	spam profiles	Google Safe Browsing API	User-based and Content- based features.	K nearest neighbors, ,Random Forest, Bayesian, Support Vector Machines	87.30%	10,782 tweets	2018
Collective Classification of Spam Campaigners on Twitter: A Hierarchical Meta-Path Based Approach [7]	TWITTER	URL-based spam	Twitter streaming API.	All features	SVM, KNN,Logistic regression , Latent Dirichlet Allocation, Decision Tree , Naive Bayes , Random Forest	feedback strategy achieves 25.6% and 46% higher F1-score and AUC	22 million tweets	2018
A Framework for Real-Time Spam Detection in Twitter [8]	TWITTER	SPAM PROFILE	Twitter streaming API.	Text based features	Support Vector Machine, Neural Network, Random Forest and Gradient Boosting. With Neural Network	91.65%	400,000 tweets	2018

Table 2 Review on Spammers

VI CONCLUSION

The technological advancement in mobile and computer and their applications opened a gate way for mischievous user account and spamming. In this paper, many articles and research publications were reviewed that deals with mischievous contents and spams. This paper focusses on four different categories of mischievous contents such as spam user accounts, fake user account, compromised user account and phishing detections. And these mischievous contents were categorized in to two main groups namely graph based and non-graph based contents. To survive in market, new researches introduce a third kind synthetic graph dataset. Finally, a literature survey is made on available online research repositories like IEEE, ACM, ScienceDirect and Springler and results are published. In the review of the communities, they are detected from benchmark databases rather then real time databases. Computational complexity will be reduced has to be in reduced for community detection. The community based nodes are evaluated on - NMI(Non mutual information),S-NMF(Symmetric Non Mutual Factor) ,ARI(Attribute Random Index),Modularity score which has to be improvised.

The paper focuses on the categories of the community deduction with the help of mind map,and a review is done for community detection methods. Once the Community is found in a given set of the dataset ,inside the community the spammers can be easily found using the metrics and various spam deduction methods used in the review.

REFERENCES

1. Malik Mateen A" hybrid approach for spam detection for Twitter" 14th International Bhurban Conference on Applied Sciences and Technology (IBCAST) DOI: 10.1109/IBCAST.2017.7868095
2. Claudia Meda, Federica Bisio, Paolo Gastaldo and Rodolfo Zunino University of Genoa "A Machine Learning Approach for Twitter Spammers Detection" 2014 International Carnahan Conference on Security Technology (ICCST) DOI: 10.1109/CCST.201
3. Nattanan Watcharenwong"Spam detection for closed Facebook groups 14th International Joint Conference on Computer Science and Software Engineering (JCSSE) DOI: 10.1109/JCSSE.2017.8025914
4. Sazzadur Rahman, Ting-Kai Huang, Harsha V. Madhyastha, and Michalis Faloutsos "Detecting Malicious Facebook Applications" in IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 24, NO. 2, APRIL 2016 pg 773-779
5. FarazAhmedMuhammadAbulaish "A generic statistical approach for spam detection in Online Social Networks" in Computer Communications "Volume 36, Issues 10–11, June 2013, Pages 1120-1129

6. Varad Vishwarupe, Mangesh Bedekar,Milind Pande, Anil Hiwale "Intelligent Twitter Spam Detection: A Hybrid Approach" in Smart Trends in Systems, Security and Sustainability pp 189-197.
7. Srishti Gupta , Abhinav Khattar,Arpit Gogia DTU Ponnurangam Kumaraguru,Tanmoy Chakraborty "Collective Classification of Spam Campaigners on Twitter: A Hierarchical Meta-Path Based Approach" in WWW '18 Proceedings of the 2018 World Wide Web Conference Pages 529-538.
8. Himank Gupta ,Mohd Saalim Jamal,Sreekanth Madisetty"A framework for real-time spam detection in Twitter" in 2018 10th International Conference on Communication Systems & Networks (COMSNETS) **DOI:** 10.1109/COMSNETS.2018.8328222
9. Ab Razak, M.F., Anuar, N.B., Salleh, R., Firdaus, A., 2016. The rise of "malware":bibliometric analysis of malware study. J. Netw. Comput. Appl. 75, 58–76.
10. Adamic, L., Adar, E., 2005. How to search a social network. Soc. Netw. 27 (3), 187–203. Adamic, L.A., Adar, E., 2003. Friends and neighbors on the web. Soc. Netw. 25 (3),211–230.
11. Al Hasan, M., Chaoji, V., Salem, S., Zaki, M., 2006. Link prediction using supervised learning.In: SDM'06: Workshop on Link Analysis, Counter-terrorism and Security.
12. Balakrishnan, V., Humaidi, N., Lloyd-Yemoh, E., 2016. Improving document relevancy using integrated language modeling techniques. Malays. J. Comput. Sci. 29 (1).
13. Bhattacharya, M., Islam, R., Abawajy, J., 2016. Evolutionary optimization: a big data perspective. J. Netw. Comput. Appl. 59, 416–426.
14. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S., 2012a. Detecting automation of twitter accounts: are you a human, bot, or cyborg? IEEE Trans. Dependable Secur. Comput. 9 (6), 811–824. http://dx.doi.org/10.1109/TDSC.2012.75.
15. Egele, M., Stringhini, G., Kruegel, C., Vigna, G., 2015. Towards Detecting compromised accounts on social networks. IEEE Trans. Dependable Secur. Comput.. http:// dx.doi.org/10.1109/TDSC.2015.2479616.
16. Ghosh, S., Viswanath, B., Kooti, F., Sharma, N.K., Korlam, G., Benevenuto, F., Gummadi, K.P., 2012. Understanding and combating link farming in the twitter social network. Proc. 21st Int. Conf. World Wide Web, 61.
17. Heidemann, J., Klier, M., Probst, F., 2012. Online social networks: a survey of a global phenomenon. Comput. Netw. 56 (18), 3866–3878 http://dx.doi.org/10.1016/j. comnet.2012.08.009.
18. Lin, P.-C., Huang, P.-M.2013. A study of effective features for detecting long-surviving Twitter spam accounts. 2013 In: Proceedings of the 15th International Conference on Advanced Communications Technology (ICACT), 841.
19. Statista, 2016. Leading social networks worldwide as of April 2016, ranked by number of active users (in millions). from⟨http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/⟩Twitter rate limit, 2015. Twitter rate limit for search/tweets REST API calls. from⟨https://dev.twitter.com/rest/public/rate-limits⟩
20. Aaron Clauset, M. E. J. Newman, and Cristopher Moore, "Finding community structure in very large networks" in Journal of Phys.Rev.E70 066111(2004).
21. Pascal Pons and Matthieu Latapy "Computing Communities in Large Networks Using Random Walks" Communities in Large Networks, JGAA, 10(2) 191–218 (2006)192.
22. L.Donetti, M.A.Munoz "Improved spectral algorithm for the detection of network communities" Journal of Phys.Rev.E05056v1(2005)
23. Usha Nandini Raghavan, Réka Albert, and Soundar Kumara "Near linear time algorithm to detect community structures in large-scale networks" Phys. Rev. E 76, 036106 – Published 11 September 2007
24. Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Go¨rke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner "On Modularity Clustering" in IEEE Transactions On Knowledge And Data Engineering, Vol. 20, No. 2, February 2008
25. Vincent D Blondel, Jean-Loup Guillaume Renaud Lambiotte and Etienne Lefebvre1 "Fast unfolding of communities in large networks" in Journal of Statiscal Mechanics doi:10.1088/1742-5468/2008/10/P10008
26. Martin Rosvall and Carl T. Bergstrom "Maps of random walks on complex networks reveal community structure" in PNAS January 29, 2008 105 (4) 1118-1123; https://doi.org/10.1073/pnas .0706851105
27. Peter Ronhovde and Zohar Nussinov "Local resolution-limit-free Potts model for community detection" in Phys. Rev. E 81, 046114 – Published 27 April 2010
28. Jianyong Wang ,yozhou chang "Parallel community detection on large networks with propinquity dynamics "in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining 2010
29. Yong-Yeol Ahn, James P. Bagrow, Sune Lehmann "Link communities reveal multiscale complexity in networks" Journal of Phys.Rev.E0903.31783(2010)
30. AaronMcaid "Detecting highly overlapping communities with model based overlapping" in 2010 International Conference on Advances in Social Networks Analysis and Mining DOI 10.1109/ASONAM.2010.77
31. Conrad Lee, Fergal Reid, Aaron McDaid, Neil Hurley "Detecting highly overlapping community structure by greedy clique expansion"in Journal of Data Analysis, Statistics and Probability
32. Steve Gregory "Finding overlapping communities in networks by label propagation" in New Journal of Physics 12 (2010) 103018 (26pp)

33. Reihaneh Rabbany Khorasgani , Jiyang Chen , Osmar R. Zaïane “Top leaders community detection approach in information networks” Proceedings of the 4th Workshop on Social Network Mining and Analysis, 2010. ISSN : 2319-7323
34. Seiji Maekawa, Koh Takeuch, Makoto Onizuka “Non-linear Attributed Graph Clustering by Symmetric NMF with PU Learning” in Journal of Advancement of Artificial Intelligence arXiv:1810.00946 2018

AUTHORS PROFILE

J.Jeyasudha is working a Assistant Professor in the Department of Software Engineering and Research Scholar in SRM Institute of Science & Technology. She has been in teaching for more than 12 years. She has taught many core subjects to the Undergraduate students of computer science engineering to build a strong technical knowledge. She is doing the research in the area of Social Engineering and Machine Learning to find the various attacks in the social media

Dr.Usha is currently working as an associate professor at the software engineering department in SRMIST.She has 11 years of teaching experience.While working in Anna University chennai she worked in research projects for Smart and Secure techniques Research Lab. Her research interest include network security, machine learning, Bio informatics. Dr.Usha published nearly 40 research articles in peer reviewed journals and international conferences. She is GATE scorer and awarded as college first rank holder in UG. She is editorial board member for the journal Progress of Electrical and Electronic Engineering. She was awarded as Outstanding Reviewer within Top 10 percentile of reviewers in Elsevier-Pattern Recognition Letters in 2017.She is reviwer of Elsevier Journal - Computer and Electrical Engineering, Elsevier Journal- Pattern Recognition Letters,Springer- Multimedia tools and Applications,IEEE Access. She has coordinated IET sponsored Workshop on Cyber Security ,National Workshop on Internet of Things , National work shop on VANET and its security IET sponsored National Conference on Big data, cloud and Security. She is a active member of IET, ISTE, Indian Science Congress. Currently she is guiding 5 Phd Students.