Journal Homepage: -www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI:10.21474/IJAR01/1753
DOI URL: http://dx.doi.org/10.21474/IJAR01/1753
ISSN NO. 2320-5407

RESEARCH ARTICLE

COMMUNITY DETECTION on SOCIAL MEDIA using GRAPH BASED APPROACH.

Aishwarya Raman and Abhishek Kanal.
Department of Computer Engineering Thadomal Shahani Engineering College.

Manuscript Info

Manuscript History

Received: 12 July 2016
Final Accepted: 22 August 2016
Published: September 2016

Key words:-
Community Detection, Social network, Applications of community detection, graph based community.

Abstract

Social media has followed an exponential graph over the past few years with incorporating features which at one time seemed impossible. The social media has had an enduring effect on the thought process of the general populace. With the diverse nature of the population which take part in the daily chatting, tagging, posting and uploading on the virtual world, the study of such coalesce of communities. This paper aims at the mining and analysis of the communities with focus on the techniques used for the detection process. We discuss four methods of detection, beginning with the node-centric moving on to group centric, then to network centric and concluding with hierarchy centric method of detection. This paper also briefly discusses the applications of community detection in varied fields.

Introduction:-

The past decade has witnessedthe rapid development of social networking sites which has empowered new ways of collaboration and communication. Social media also helps reshape business models, sway opinions and emotions, and opens up numerous possibilities to study human interaction and collective behavior in an unparalleled scale[2]. Hence, study of social network is of great importance in sociology, biology and computer science. Social network analysis is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. Social network analysis provides both a visual and a mathematical analysis of human relationships. A valuable tool in the analysis of large complex networks is community detection.

Community Detection:-

Community is formed by individuals such that those within a group interact with each other more frequently than with those outside the group[1]. There are two types of communities in social networks-

Explicit groups which are formed as a result of conscious human decision.
Implicit groups which emerge from interactions and activities of users.
Often communities are defined with respect to a graph, which consists of set of objects called vertices (V) and their relations called as edges (E). Therefore, according to computer science, ***community detection*** is identifying a group of vertices that are more densely connected to each other than the rest of the network [1]. Figure below shows a network with three communities.

Corresponding Author:-Aishwarya Raman.
Address:-Department of Computer Engineering Thadomal Shahani Engineering College.

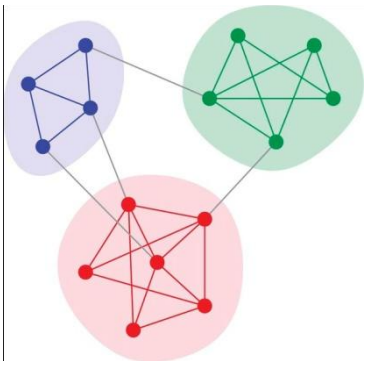


Figure 1:- Network with three communities.

Methods of Community Detection:-

Node-Centric Community Detection:-

Node-Centric community detection is commonly used in traditional social network analysis. In this type of community detection, each node in a group (community) satisfies certain properties.

Complete Mutuality- To satisfy this criterion cliques in a graph are found. A clique is an ideal cohesive subgroup. It is a maximum complete sub graph in which all nodes are adjacent to each other [2].To find maximum clique in large network recursive pruning procedure is applied. For a clique of size k, each node in the clique should maintain at least degree k − 1. Hence, those nodes with degree less than k − 1 cannot be included in the maximum clique, thus can be pruned [2]. The procedure is as follows

- A sub-network is sampled from the given network. A clique in the sub-network can be found in a greedy manner, e.g., expanding a clique by adding an adjacent node with the highest degree.
- The maximum clique found on the sub-network (say, it contains k nodes) serves as the lower bound for pruning. That is, the maximum clique in the original network should contain at least k members. Hence, in order to find a clique of size larger than k, the nodes with degree less than or equal to k − 1, in conjunction with their connections can be removed from future consideration. As social media networks follow a power law distribution for node degrees, i.e., the majority of nodes have a low degree, this pruning strategy can reduce the network size significantly.
- This process is repeated until the original network is shrunk into a reasonable size and the maximum clique can either be identified directly, or have already been identified in one of the sub-networks.[2]

Figure below shows a sub network.

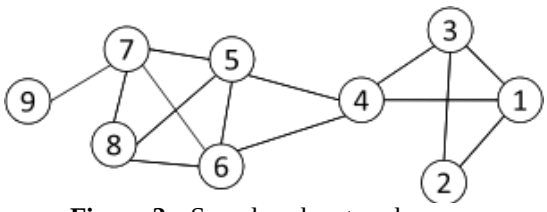


Figure 2:- Sample sub network.

The maximum clique to the given network is found as follows
Suppose we sample the sub-network with nodes {1-9} and find a clique {1, 2, 3} of size 3
In order to find a clique >3, remove all nodes with degree <=3-1=2

- ❖ Remove nodes 2 and 9
- ❖ Remove nodes 1 and 3
- ❖ Remove node 4

The resulting sub graph is

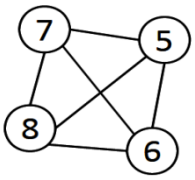


Figure 3:- Resultant sub graph.

Reachability- In node-centric community detection, reachability between two nodes is considered. Reachability can be defined using *geodesic distance*. *Geodesic* is the shortest path between any two nodes. *Geodesicdistance* is the number of hops in a geodesic between two nodes. *Geodesic diameter* is the maximal geodesic distance for any 2 nodes in a network [2]. Any node in a community should be reachable in k hops. Based on this criterion there are two types of substructures, which can be found.

a. k-clique is a maximal sub graph in which the largest geodesic distance between any two nodes is no greater than k. That is, d (v_i, v_j) ≤ k ∀v_i, v_j∈V_s where V_s is the set of nodes in the sub graph. Note that the geodesic distance is defined on the original network. Thus, the geodesic is not necessarily included in the group structure. Therefore, a k-clique may have a diameter greater than k.

b. k-club restricts the geodesic distance within the group to be no greater than k. It is a maximal substructure of diameter k.[2]

For the sub network below, 3-clique and 3-club are found.

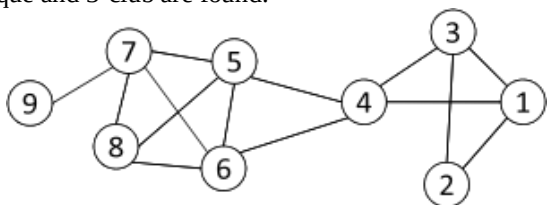


Figure 4:- Sub Network with identified cliques.

3-clique{1,2,3,4,5,6,7,8}, {4,5,6,7,8,9}
3-club:{1,2,3,4,5,6},{1,3,4,5,6,7,8}, {4,5,6,7,8,9}

Group centric community detection:-

It considers the connections within a group as a whole. Certain nodes in the group can have low connectivity, but the overall group should satisfy certain criteria. An example of group-centric community detection is finding density based groups. A sub graphG_s (V_s, E_s) is γ -dense (also called a quasi-clique [3] if

$$\frac{2|E_s|}{|V_s|(|V_s|-1)} \geq \gamma$$

Network centric community detection:-

Network-centric criterion needs to consider the connections within a network globally. Network-centric community detection partitions the whole network into several disjoint sets. There are various approaches to this type of community detection.

Vertex similarity- Vertex similarity is defined in terms of the similarity of their social circles, e.g., the number of friends two share in common. Similarity measures used in practical networks include Jaccard similarity [4] and cosine similarity [5].

Jaccard Similarity

$$Jaccard(v_i, v_j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|}$$

Cosine Similarity

$$Cosine(v_i, v_j) = \frac{|N_i \cap N_j|}{\sqrt{|N_i| \cdot |N_j|}}$$

For the given graph Jaccard and Cosine similarity are

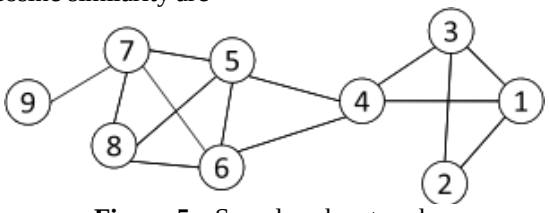


Figure 5:- Sample sub network

$$Jaccard(4,6) = \frac{|\{5\}|}{|\{1,3,4,5,6,7,8\}|} = \frac{1}{7}$$

$$cosine(4,6) = \frac{1}{\sqrt{4 \cdot 4}} = \frac{1}{4}$$

Latent space models- A latent space model maps nodes in a network into a low-dimensional Euclidean space such that the proximity between nodes based on network connectivity are kept in the new space [6][7], then the nodes are clustered in the low-dimensional space using methods like k-means[8]. One representative approach is multi-dimensional scaling (MDS) [9]. Typically, MDS requires the input of a proximity matrix$P \in R^{n \times n}$, with each entry P_{ij} denoting the distance between a pair of nodes i and j in the network. $S \in R^{n \times l}$ denote the coordinates of nodes in the l-dimensional space such that S is column orthogonal. It can be shown that

$$SS^T \approx -\frac{1}{2}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(P \circ P)(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T) = \widetilde{P}$$

where I istheidentitymatrix,1 an n-dimensional column vector with each entry being 1,and ◦ the element-wise matrix multiplication. It follows that S can be obtained via minimizing the discrepancy

$$\min \|SS^T - \widetilde{P}\|_F^2$$

Suppose V contains the top l eigenvectors of P with largest Eigenvalues, Λ is a diagonal matrix of top l eigenvalues $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_l)$.The optimal S is $S = V\Lambda^{1/2}$[2]. The classical k-means algorithm can be applied to S to find community partitions.

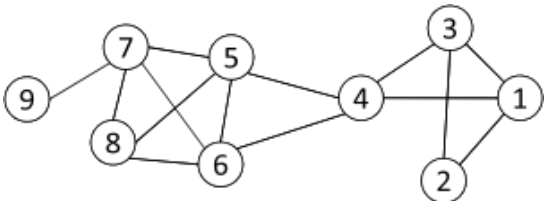


Figure 6:- Sample sub network.

$$P = \begin{bmatrix} 0 & 1 & 1 & 1 & 2 & 2 & 3 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 & 3 & 4 & 4 & 5 \\ 1 & 1 & 0 & 1 & 2 & 2 & 3 & 3 & 4 \\ 1 & 2 & 1 & 0 & 1 & 1 & 2 & 2 & 3 \\ 2 & 3 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \\ 2 & 3 & 2 & 1 & 1 & 0 & 1 & 1 & 2 \\ 3 & 4 & 3 & 2 & 1 & 1 & 0 & 1 & 1 \\ 3 & 4 & 3 & 2 & 1 & 1 & 1 & 0 & 2 \\ 4 & 5 & 4 & 3 & 2 & 2 & 1 & 2 & 0 \end{bmatrix}$$

$$\widetilde{P} = \begin{bmatrix} 2.46 & 3.96 & 1.96 & 0.85 & -0.65 & -0.65 & -2.21 & -2.04 & -3.65 \\ 3.96 & 6.46 & 3.96 & 1.35 & -1.15 & -1.15 & -3.71 & -3.54 & -6.15 \\ 1.96 & 3.96 & 2.46 & 0.85 & -0.65 & -0.65 & -2.21 & -2.04 & -3.65 \\ 0.85 & 1.35 & 0.85 & 0.23 & -0.27 & -0.27 & -0.82 & -0.65 & -1.27 \\ -0.65 & -1.15 & -0.65 & -0.27 & 0.23 & -0.27 & 0.68 & 0.85 & 1.23 \\ -0.65 & -1.15 & -0.65 & -0.27 & -0.27 & 0.23 & 0.68 & 0.85 & 1.23 \\ -2.21 & -3.71 & -2.21 & -0.82 & 0.68 & 0.68 & 2.12 & 1.79 & 3.68 \\ -2.04 & -3.54 & -2.04 & -0.65 & 0.85 & 0.85 & 1.79 & 2.46 & 2.35 \\ -3.65 & -6.15 & -3.65 & -1.27 & 1.23 & 1.23 & 3.68 & 2.35 & 6.23 \end{bmatrix}$$

$$V = \begin{bmatrix} -0.33 & 0.05 \\ -0.55 & 0.14 \\ -0.33 & 0.05 \\ -0.11 & -0.01 \\ 0.10 & -0.06 \\ 0.10 & -0.06 \\ 0.32 & 0.11 \\ 0.28 & -0.79 \\ 0.52 & 0.58 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 21.56 & 0 \\ 0 & 1.46 \end{bmatrix}, \quad S = V\Lambda^{1/2} = \begin{bmatrix} -1.51 & 0.06 \\ -2.56 & 0.17 \\ -1.51 & 0.06 \\ -0.53 & -0.01 \\ 0.47 & -0.08 \\ 0.47 & -0.08 \\ 1.47 & 0.14 \\ 1.29 & -0.95 \\ 2.42 & 0.70 \end{bmatrix}$$

k-meanscanbeappliedto S inordertoobtaindisjointpartitionsofthenetwork.At the end, we obtain two clusters {1,2,3,4},{5,6,7,8,9}, which can be represented as a partition matrix[2].

Block model approximation- Block models approximate a given network by a block structure. Each block represents one community. Therefore, we approximate a given adjacency matrix A as follows.

$$A \approx S\Sigma S^T$$

where $S \in \{0,1\}^{n \times k}$ is the block indicator matrix with S_{ij} =1 if node i belongs to the j-thblock,Σa$k \times k$ matrix indicating the block (group) interaction density, and k the number of blocks. A natural objective is to minimize the following[2]

$$\min ||A - S\Sigma S^T||_F^2$$

For the given graph, the top two Eigen vectors of the adjacency matrix are

$$S = \begin{bmatrix} 0.20 & -0.52 \\ 0.11 & -0.43 \\ 0.20 & -0.52 \\ 0.38 & -0.30 \\ 0.47 & 0.15 \\ 0.47 & 0.15 \\ 0.41 & 0.28 \\ 0.38 & 0.24 \\ 0.12 & 0.11 \end{bmatrix}, \Sigma = \begin{bmatrix} 3.5 & 0 \\ 0 & 2.4 \end{bmatrix}.$$

As indicated by the sign of the second column of S, nodes {1,2,3,4} form a community, and {5,6,7,8,9}is another community, which can be obtained by a k-means clustering applied to S.

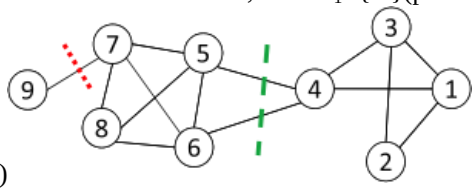
Spectral Clustering- Spectral clustering is derived from the problem of graph partition. Graph partition aims to find out a partition such that the cut (the total number of edges between two disjoint sets of nodes) is minimized [10]. Two commonly used variants in community detection are ratio and *normalized* cut. Le $t\pi = (C_1,C_2,\cdots,C_k)$ be a graph partition such that $C_i \cap C_j = \varphi$ and $\cup^k_{i=1} C_i = V$.Theratiocutandthenormalizedcut are defined as:

$$\text{Ratio Cut}(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i,\bar{C}_i)}{|C_i|},$$

$$\text{Normalized Cut}(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i,\bar{C}_i)}{vol(C_i)}$$

where $\bar{}$ Ci is the complement of Ci, and
vol (Ci) $=\Sigma_{v\in Ci}\, d_v$.

Suppose we partition the network below into two communities, with C_1={ 9}(partition in red) and



C_2={1,2,3,4,5,6,7,8} (partition in green)

Figure 7:- Sample sub network with ratio cuts.

For partition in red(π_1)

$$\text{Ratio Cut}(\pi_1) = \frac{1}{2}\left(\frac{1}{1}+\frac{1}{8}\right) = 9/16 = 0.56$$

$$\text{Normalized Cut}(\pi_1) = \frac{1}{2}\left(\frac{1}{1}+\frac{1}{27}\right) = 14/27 = 0.52$$

For partition in green(π_2)

$$\text{Ratio Cut}(\pi_2) = \frac{1}{2}\left(\frac{2}{4}+\frac{2}{5}\right) = 9/20 = 0.45 < \text{Ratio Cut}(\pi_1)$$

$$\text{Normalized Cut}(\pi_2) = \frac{1}{2}\left(\frac{2}{12}+\frac{2}{16}\right) = 7/48 = 0.15 < \text{Normalized Cut}(\pi_1)$$

Modularity Maximization:-

Modularity is proposed specifically to measure the strength of a community partition for real-world networks by taking into account the degree distribution of nodes[11]. Given a network of n nodes and m edges, the expected number of edges between nodes v_i and v_j is $d_id_j/2m$, where d_i and d_j are the degrees of node v_i and v_j, respectively. Considering one edge from node vi connecting to all nodes in the network randomly, it lands at node v_j with probability $d_j/2m$. As there are di such edges,the expected number of connections between the two are $d_id_j/2m$[2]. For the graph below the expected number of edges between nodes 1 and 2 is 3*2/ (2*14) = 3/14.

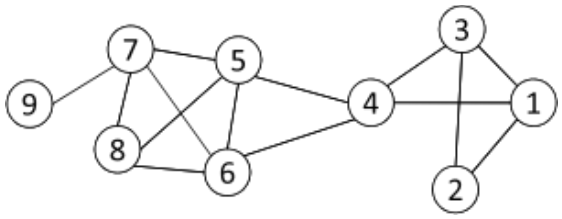


Figure 8:- Sample sub network

So $A_{ij} - d_i d_j/2m$ measures how far the true network interaction between nodes i and j (A_{ij}) deviates from the expected random connections. Given a group of nodes C, the strength of community effect is defined as

$$\sum_{i \in C, j \in C} A_{ij} - d_i d_j / 2m$$

Modularity is defined as

$$Q = \frac{1}{2m} \sum_{\ell=1}^{k} \sum_{i \in C_\ell, j \in C_\ell} (A_{ij} - d_i d_j / 2m)$$

where the coefficient 1/2m is introduced to normalize the value between -1 and 1. Modularity calibratesthequalityofcommunitypartitionsthuscanbeusedasanobjectivemeasuretomaximize. Equivalently,

$$\boldsymbol{B = A - \mathbf{d}\mathbf{d}^T/2m}$$

Modularity maximization can be reformulated as

$$\max Q = \frac{1}{2m} Tr(S^T B S) \quad s.t.\ S^T S = I_k$$

With a spectral relaxation to allow S to be continuous, the optimal S can be computed as the top keigenvectors of the modularity matrix B [11] with the maximum eigenvalues.

$$B = \begin{bmatrix} -0.32 & 0.79 & 0.68 & 0.57 & -0.43 & -0.43 & -0.43 & -0.32 & -0.11 \\ 0.79 & -0.14 & 0.79 & -0.29 & -0.29 & -0.29 & -0.29 & -0.21 & -0.07 \\ 0.68 & 0.79 & -0.32 & 0.57 & -0.43 & -0.43 & -0.43 & -0.32 & -0.11 \\ 0.57 & -0.29 & 0.57 & -0.57 & 0.43 & 0.43 & -0.57 & -0.43 & -0.14 \\ -0.43 & -0.29 & -0.43 & 0.43 & -0.57 & 0.43 & 0.43 & 0.57 & -0.14 \\ -0.43 & -0.29 & -0.43 & 0.43 & 0.43 & -0.57 & 0.43 & 0.57 & -0.14 \\ -0.43 & -0.29 & -0.43 & -0.57 & 0.43 & 0.43 & -0.57 & 0.57 & 0.86 \\ -0.32 & -0.21 & -0.32 & -0.43 & 0.57 & 0.57 & 0.57 & -0.32 & -0.11 \\ -0.11 & -0.07 & -0.11 & -0.14 & -0.14 & -0.14 & 0.86 & -0.11 & -0.04 \end{bmatrix}$$

Its top two maximum eigenvectors are

0.4384	-0.2709
0.3809	0.2671
0.4384	-0.2709
0.1716	0.6063
-0.2861	-0.3487
-0.2861	-0.3487
-0.3754	0.3355
-0.3421	0.1855
-0.1396	-0.1552

Hierarchy centric community detection:-

Another line of community detection research is to build a hierarchical structure of communities based on network topology. This facilitates the examination of communities at different granularity. There are mainly two types of hierarchical clustering: divisive, and agglomerative

Divisive:-

One particular divisive clustering algorithm is to recursively remove the "weakest" tie in a network until the network is separated into two or more components. The general principle is as follows:

- ❖ At each iteration, find out the edge with least strength. This kind of edge is most likely to be a tie connecting two communities.
- ❖ Remove the edge and then update the strength of links.
- ❖ Once a network is decomposed into two connected components, each component is considered a community.

The iterative process above can be applied to each community to find sub communities.

Newman and Girvan proposed a method to find weak ties using *edge betweeness. Edge betweenness*is defined to be the number of shortest paths that pass along one edge (Brandes, 2001). .TheNewman-Girvan algorithm suggests progressively removing edges with the highest betweenness. It will gradually disconnect the network,naturally leading to a hierarchical structure. [2]

Edge betweeness of the figure below id shown in the table.

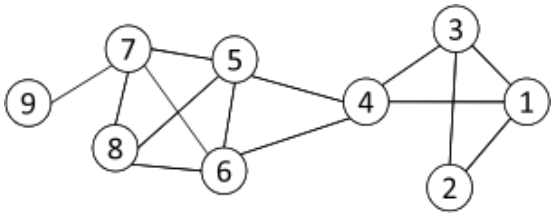


Figure 9:- Sample sub network.

	1	2	3	4	5	6	7	8	9
1	0	4	1	9	0	0	0	0	0
2	4	0	4	0	0	0	0	0	0
3	1	4	0	9	0	0	0	0	0
4	9	0	9	0	10	10	0	0	0
5	0	0	0	10	0	1	6	3	0
6	0	0	0	10	1	0	6	3	0
7	0	0	0	0	6	6	0	2	8
8	0	0	0	0	3	3	2	0	0
9	0	0	0	0	0	0	8	0	0

Figure 10:- Edge betweeness of above subgraph.

The Newman and Girvan algorithm is applied to the sample sub network. The process is

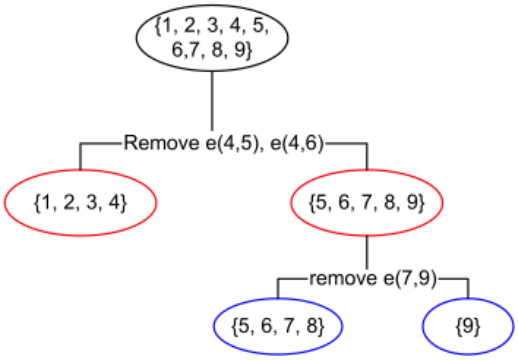


Figure11:- Newman and Girvan Algorithm Flowgraph.

Agglomerative- Agglomerative clustering begins with base communities and merges them successively into larger communities following certain criterion. One such criterion is modularity (Clauset et al., 2004). Two communities are merged if doing so results in the largest increase of overall modularity.

Figure shows the resultant dendrogram based on agglomerative hierarchical clustering applied to the sample network. Nodes 7 and 9 aremergedfirst,and then 1 and 2, and so on. Finally, we obtain two communities at the top{1,2,3,4}and{5,6,7,8,9}.[2]

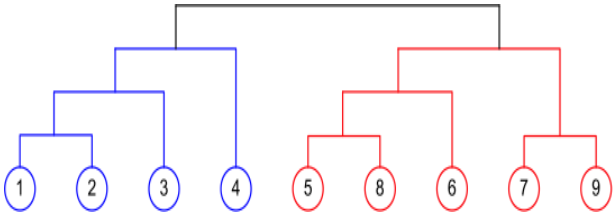


Figure 12:- Resultant Dendrogram.

Applications of community detection:-

Detection of suspicious events in social media:-

Social network analysis can be used to increase the knowledge about the customers' behavior, mostly in relation to the customers' connections and how they create communities according to their call and text messages. By performing community detection, it is possible to recognize groups of customers which unexpected behavior in terms of usage and also in regard to types of social structures. Outliers groups might be pointed out as suspicious communities in terms of fraud events [12].

Recommendation systems:-

Community detection can be used to build recommender systems, which recommends the most suitable products to the customers by predicting their interest. When focusing on the problem of recommending items to a user (i.e. a customer of an e-store), the underlying transaction data can be seen as a complex network (specifically, a bipartite network): inside this structure, information about customer tastes is codified and can be of good use for future suggestions[13].

Link prediction:-

Community detection in complex networks can be used for link prediction between two actors. Link prediction evaluates the possibility of existence of future links between vertices by observing vertices and links attributes in the network. Link prediction is used to detect missing and fake links and predicts future existence of the links with the development of network [14].

Detection of terrorist groups:-

With the increasing popularity of social media over the last few years, terrorist groups have flocked to the popular web sites to spread their message and recruit new members. As terrorist groups establish a presence in these social networks, they do not rely on direct connections to influence sympathetic individuals. Instead, they leverage "friend of a friend" relationships where existing members or sympathizers bridge the gap between potential recruits and terrorist leadership or influencers. These terrorist social networks in social media can be uncovered and mapped, providing an opportunity to apply social network analysis algorithms. Leveraging these algorithms, the main influencers can be identified along with the individuals bridging the gap between the sympathizers and influencers [15].]

Anomaly detection in social media:-

Anomalies in online social networks can signify irregular, and often illegal behavior. Detection of such anomalies has been used to identify malicious individuals, including spammers, sexual predators, and online fraudsters. The detection of anomalies in online social networks is composed of two sub-processes; the selection and calculation of network features, and the classification of observations from this feature space [16].

Conclusion:-

In this paper, we discussed the concept of graph-based community and community detection. Methods of community detection was explained using appropriate examples - Node centric community detection, group centric community detection, network centric community detection and hierarchy centric community detection. We also discussed applications of community detection- detection of suspicious events, recommendation systems, link prediction, detection of terrorist groups in social network and anomaly detection.

References:-

1. Orgnet," http://www.orgnet.com/sna.html"
2. Morgan & Claypool, "Community detection and mining in social media
3. Abello et al., 2002
4. Gibson et al., 2005
5. Hopcroft et al., 2003
6. Hoff et al., 2002
7. Handcock et al., 2007
8. Tan et al., 2005
9. Borg and Groenen, 2005
10. Luxburg, 2007
11. Newman, 2006
12. Carlos André Reis Pinheiro, "Community Detection to Identify Fraud Events in Telecommunications Networks"
13. MassimilianoZanin, Pedro Cano, "Complex Networks in Recommendation Systems".
14. Fei Tan, Yongxiang Xia,* and Boyao Zhu, "Link Prediction in Complex Networks: A Mutual Information Perspective"
15. Todd Waskiewicz (Air Force Research Laboratory, AFRL/RIEA 525 Brooks Road, Rome, NY 13441-4505), "Friend of a Friend Influence in Terrorist Social Networks"
16. David Savage, Xiuzhen Zhang, Xinghuo Yu, Pauline Chou[a],Qingmai Wang, "Anomaly detection in social media"