IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

COMMUNITY DETECTION IN SOCIAL NETWORKS USING UNNORMALIZED SPECTRAL CLUSTERING COMBINED WITH KNN ALGORITHM

Kurapati Sravanthi

CSE Department,

University College of Engineering (Kakatiya University), Kothagudem, Telangana, INDIA.

Abstract: Social network analysis has gained much attention in recent times. A graph can be used to depict social networks. In the analysis of social networks, every individual is denoted as a node, and the connections between them are represented as edges. Identifying communities and representing the interactions between entities and individuals in real-world network graphs is a difficult task. There are numerous established methods for locating the linked nodes that eventually result in the discovery of communities. This paper presents a novel method for community detection in social networks by integrating unnormalized spectral clustering with the k-nearest neighbors (KNN) algorithm. The approach leverages the strengths of spectral clustering for global structure analysis and KNN for local neighborhood refinement. The primary objective of the algorithm proposed in this study is to identify and eliminate any noisy nodes from the identified communities, hence enhancing the quality of the identified communities. Experimental results on synthetic and real-world datasets demonstrate the method's effectiveness in accurately identifying community structures.

Index Terms **- Community Detection, Social Networks, Spectral Clustering, K-Nearest Neighbors (KNN).**

I. INTRODUCTION

In recent years, the analysis of social networks [1] has gained significant attention due to its wide-ranging applications in various fields such as sociology, marketing, and information dissemination. Some of the real-world networks include network of co-authorship [2], biological networks that includes neural networks [3], the World Wide Web (WWW) (e.g., a network of hyperlinks of web pages), network of friendship, food webs [4], technological networks (e.g., Internet), metabolic networks [5], social networks, and even political elections. Graphs are commonly used to represent social networks, with nodes standing in for individuals and edges for the connections between them [6]. One important task that can disclose the underlying structure and function of the network is identifying communities within these graphs. Communities, also known as clusters, are collections of nodes that exhibit higher levels of connectivity among themselves compared to the rest of the network. These groups are indicative of entities that have common behaviors, interests, or functions. A schematic of a basic network with a community structure is presented in Fig. 1.

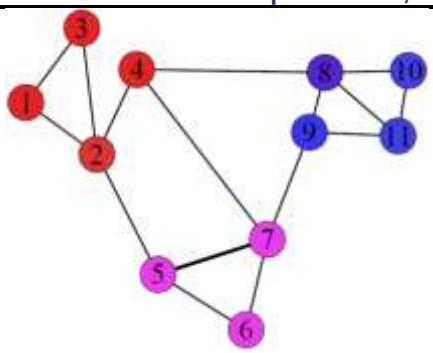

Fig. 1. A schematic diagram showing a social network with three community structures. (Drawn by using social network visualization tool Gephi)

As stated by pizzuti (2008), "a social network can be modeled as a graph G = (V, E) where V is a set of objects, called vertices or nodes, and E is a set of edges, called links, that connect two elements of V" (p.l082) [7]. Identifying groups of subsets of nodes where the density of edges inside the subset is high and the density of edges between subsets is low is known as the community discovery problem in social networks [8]. The illustrating features of social networks can be understood more clearly by such community identification and exploit them more effectively. For example, we can find a set of web pages on related topics by identifying a community of web pages that connect two or more web pages in the same community; with the help of this, the search engines and portals can narrow down their search by searching topically-related subsets of web pages [9].

The organization of the paper is as follows. In Section II, literature concerning related work is reviewed. In Section III, we give a formal problem statement and proposed algorithm description. Experimental results are presented in Section IV. In Section V we describe the conclusions along with future work.

II. RELATED WORK

Communities have a long history in social sciences. In recent times different algorithms have been proposed by different disciplines such as mathematics, machine learning, statistics, data mining, for detecting communities in social and other complex networks [10]. In the following, literature about some of the well-known algorithms is reviewed.

Girvan and Newman first stated the idea of community structure in social networks [11]. According to Girvan and Newman many real networks contain sub graphs of nodes appearing as a group such that the density of internal connections between nodes of sub graph is larger than the connections with the remaining nodes in the network. A divisive hierarchical method is proposed by them in [11]. In this technique edges with high betweenness are removed one after the other such that the graph is divided into clusters hierarchically.

In [12], Hopcroft presented an agglomerative hierarchical clustering method to identify stable or natural communities in large linked networks. According to Hopcroft (2003), "a cluster is assumed as a natural cluster if it appears in the clustering process when a given percentage of links are removed".

In [13], Raju et. al. proposed a Cohesion Index based Label Propagation Algorithm (CILPA) to identify community structure in complex social networks. The algorithm brings in two new functions, called Cohesion Similarity (CoSim) and Cohesion Index (CI). The cohesion index function measures cohesiveness of nodes and similarity with neighbor nodes is measured using cohesion similarity. Cohesion index as the base, they proposed a new label propagation algorithm with precise node update sequence and node priority.

In [14], Raju et. al. proposed a clustering coefficient-based label propagation algorithm (CCLPA) for unfolding communities in complex networks. They devised the CCLPA algorithm to address the randomness issue of label propagation algorithm. The algorithm defines a function, clustering coefficient, to measure the neighborhood connectivity between nodes quantitatively without any contact with the user. Based on the clustering coefficient, they presented a new label propagation algorithm with explicit node update sequence to uncover communities in complex networks.

In social networks, to identify communities, the authors of [15] presented a spectral clustering method. In this method, core members are used by the authors for extracting communities. The authors used page rank algorithm for detection of communities for complete use of network features and proved that their method is better in terms of time and accuracy.

III. PRELIMINARIES

In this section we briefly discuss spectral clustering method first, then define community finding problem in a network, and then focus on describing the proposed algorithm.

A. Spectral Clustering

In clustering of data points, the spectral clustering algorithm performs dimensionality reduction on eigen values (spectrum) obtained from the similarity matrix. It then performs clustering on fewer dimensions. The input to the algorithm is the similarity matrix. The similarity matrix contains the relative similarity between each pair of points in the given dataset. This relative similarity is assessed quantitatively.

Given a set, D, of data points, the similarity matrix, may be defined as a symmetric matrix W, where W_{ij}> = O represents a measure of the similarity between data points with indexes i and j. The unnormalized graph Laplacian matrix is defined as

$$L = D - W.$$

The rest of the algorithm is described in part C of this section.

B. Problem Statement

In networks, the aim of community detection problem is defined as to find a partition $C = \{ c_1, c_2, \ldots, c_k \}$ of a simple graph G = (V, E), where $\forall_i$, $c_i \subseteq V$ and $\forall_{i,j}$, $c_i \cap c_j = \Phi$. Each c_i $i = 1, \ldots, k$ is a sub-graph containing a group of vertices of G. This subgraph, c_i is known as community such that the intra-cluster density of edges within the sub-graph is high and inter-cluster density of edges is low.

In the following, we suppose G is an undirected graph without multiple edges. Letters *i, j* indicate nodes; *e(i, j)* represents an edge connecting the nodes *i* and *j*. The adjacency matrix representation of graph is used to work out graph problems. The adjacency matrix A of a graph G is represented by an $n \times n$ matrix containing 0's and 1's, $A = (a_{ij})_{n \times n}$ where $a_{ij} = 1$ if there is an edge between the vertices *i* and *j*; otherwise, $a_{ij} = 0$. The adjacency matrix for an undirected graph is symmetric. A sample social network [16] is shown in fig. 2 and its adjacency matrix is shown in Table. I.

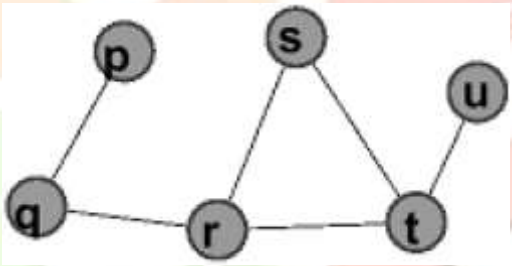

Fig. 2. A Sample Social Network

TABLE.1. ADJACENCY MATRIX OF SOCIAL NETWORK SHOWN IN FIG.2.

Node	p	q	r	s	t	u
p	0	1	0	0	0	0
q	1	0	1	0	0	0
r	0	1	0	1	1	0
s	0	0	1	0	1	0
t	0	0	1	1	0	1
u	0	0	0	0	1	0

C. Proposed Algorithm

Following is pseudo code of the algorithm Unnormalized Spectral Clustering combined with KNN.

Unnormalized Spectral Clustering combined with KNN Algorithm

Input: Similarity matrix $S \in R^{nxn}$, number k of clusters to be constructed.

- Construct a similarity graph. Let W be its weighted adjacency matrix.
- Compute the unnormalized Laplacian L.
- Compute the first k eigen vectors $u_1, u_2, \ldots, u_k$ of L.
- Let U ϵ R^{nxk} be the matrix containing the vectors $u_1, u_2, \ldots, u_k$ as columns.
- For $i = 1, ..., n$, let $y_i \in R^k$ be the vector corresponding to the i^{th} row of U.
- Cluster the points $(y_i)_{i=1,...,n}$ in R^k with the KNN algorithm into clusters C1, • • • , Ck.

Output: Clusters $P_1, ... ,P_k$ with $Pi = \{ j / y_j \in C_i \}$.

The pseudo code of the K-Nearest Neighbor (KNN) algorithm:

Consider k as the desired number of nearest neighbors and $S:=p_1,..., p_n$ be the set of training samples in the form $p_1=(x_i,c_i)$, where x_i is the d-dimensional feature vector of the point p_i and c_i is the class that p_i belongs to.

For each p'=(x', c')

- Compute the distance $d(x',x_i)$ between p' and all p_i belonging to S.
- Sort all points p_i according to the key $d(x', x_i)$
- Select the first k points from the sorted list, those are the k closest training samples to p'
- Assign a class to p' based on majority vote: $c'=argmax_y\sum_{(xi,ci)}$ belonging to S, $I(y=c_i)$.

End

IV. EXPERIMENTAL RESULTS

In this section, results are presented for various applications to which our unnormalized spectral clustering combined with KNN is applied. The algorithm is applied on Zachary Karate Club [17] and American College Football [18] network datasets. In each of these cases we find that our algorithm detects the community structures in a reliable manner. Table II gives some statistics of these datasets [16].

A. Zachary Karate Club dataset

In the early 1970s, at an American university, Wayne Zachary studied the members of a karate club for two years and recorded their social interactions. Based on their social interactions, he built a network dataset with 34 vertices and 78 edges. In this dataset, the students were represented as vertices and two students are linked by an edge if they are good friends. By chance, a dispute arose during the course of his study between the club's administrator and the karate teacher. As a result, the club splits into two smaller communities with the administrator and the teacher being as the central persons accordingly. The original division of the club into 2 communities is shown in Fig 3.

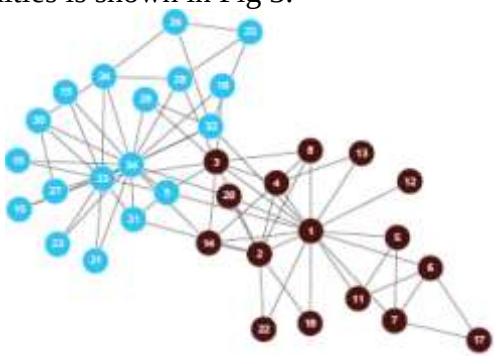

Fig. 3 Zachary Karate Club Network with two communities

TABLE II. STATISTICS OF EXPERIMENTAL DAT ASETS

Dataset	|V|	|E|	<k>	C	D
Zachary Karate Club	34	78	4.58	0.56	5
American College Football Network	115	616	10.6	0.40	4
Bottlenose Dolphin Network	62	159	5.13	0.26	8

<k> - average degree of the dataset
C - Clustering Coefficient of the dataset
D - Diameter of the dataset

Figure 4 shows the results of our approach applied on Zachary Karate Club Network. The algorithm is executed for 10 different runs and presented the average results of these 10 runs. The results are compared with the benchmark algorithm.

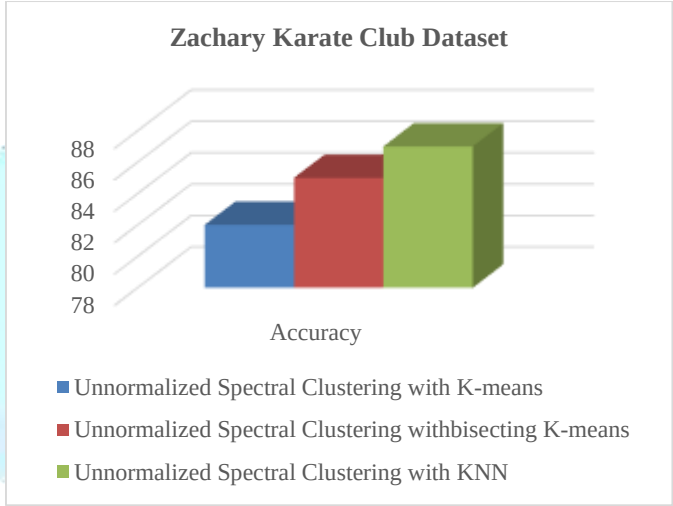

Fig. 4. Results of three algorithms on Zachary Karate Club Dataset.

B. American College Football

The American College Football network dataset was developed from the United States college football games. The schedule of games between Division IA colleges during the season Fall 2000 is represented by this network. Teams are represented by vertices in the network and the regular season games between two teams are represented by edges. The total number of vertices in this dataset is 115 and the number of edges is 616. The teams are divided into conferences. Each team in each conference, on an average, played 4 matches with teams of same conference and 7 matches with teams of other conferences. Figure 5 shows the actual community structure of this dataset before applying the algorithm.

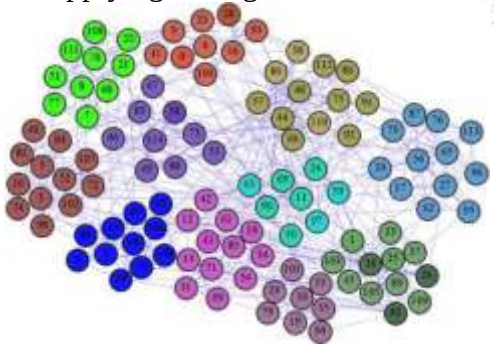

Fig. 5. American College Football Network Dataset with 11 communities

When the proposed algorithm unnormalized spectral clustering combined with KNN is applied on this dataset we achieved good results. The algorithm is executed ten times and the results obtained are displayed in Table III along with the conference name and its size, and the number of times the actual grouping is identified successfully by the algorithm. On an average, the number of teams which are misplaced is also displayed in the table whenever an improper community is found. The same kind of information about the benchmark algorithm also appears in the table. The letter "Y" in table III indicates that the group detected is the actual group, the integer number displayed, instead, tells that the number of teams which are wrongly assigned to other communities by the algorithm.

Table III also displays that, over the ten runs, the Unnormalized Spectral Clustering with Bisecting k-means incorrectly grouped four conferences teams, namely Conference USA, Western Athletic, MidAmerican and Sun belt. Even the benchmark algorithm, and also Girvan and Newman algorithm failed in such cases. Since there is no much difference in scheduling these games, the failure is due to poor maintenance of conference structure in these cases. The results obtained reveal the capability of unnormalized spectral clustering combined with KNN to deal with community detection in networks, effectively.

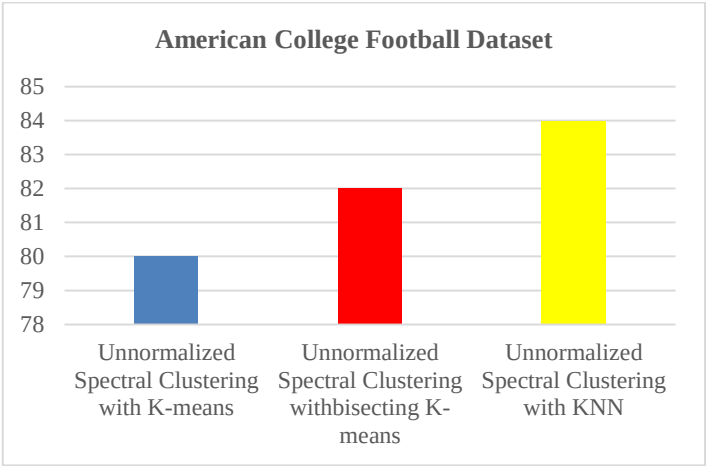

Fig. 6. Results of three algorithms on American College Football Dataset.

TABLE III. RESULTS ON AMERICAN COLLEGE FOOTBAL DATASET

Conference	No.of Teams	Benchmark Algorithm		Proposed Algorithm		
		Number of Correct grouping	Average num of misclassified teams	Number of Correct grouping	Average no. of misclassified teams	
Atlantic Coast	12	10/10	–	10/10	–	Y
Sun Belt	8	0/10	3	3/10	1	3
Big East	8	6/10	3	7/10	1	Y
Pacific Ten	10	9/10	1	9/10	1	Y
Big Ten	11	5/10	1.5	7/10	1	Y
Western Athletic	9	0/10	4	4/10	1.5	2
Big Twelve	12	10/10	–	10/10	–	Y
Southeastern	12	6/10	3	8/10	1	Y
Conference USA	12	2/10	5.5	4/10	2.6	2
Mid-American	12	7/10	3.5	9/10	1.5	5
Mountain West	9	5/10	2	6/10	1	Y

V. CONCLUSION AND FUTURE WORK

In this paper we have reported an experimental study of unnormalized spectral clustering combined with KNN algorithm to identify the community organization of underlying network datasets. The proposed algorithm is tested on two real world social networks: Zachary'S Karate Club network and American College Football teams' network. Experimental results confirm the effectiveness of the proposed approach.

After observing the results of the experiments, it is clearly seen that unnormalized spectral clustering combined with KNN algorithm gets the best results on all data sets compared to the benchmark algorithm.

Detecting automatically the number of clusters present in the network, and thereafter, analyzing such underlying community structure is an interesting future research direction to be carried out.

REFERENCES

[1] Raju, E. and Sravanthi, K. 2012. Analysis of Social Networks Using the Techniques of Web Mining. International Journal of Advanced Research in Computer Science and Software Engineering, 2(10): 443-450.

[2] Nascimento, M. A. and Jorg, S. and Pound, J. 2003. Analysis of sigmod's co-authorship graph. SIGMOD Record, 32(2):57-58.

[3] Watts, D. J. and Strogatz, S. H.1998. Collective dynamics of small-world networks. Nature, 393: 440-442.

[4] Williams, R. J. and Martinez, N. D. 2000. Simple rules yield complex food webs. Nature, 404: 180-183.

[5] Jeong, H., B. Tombor, R,. Albert, Z.N., Oltvai and Barabasi, A.2000. The large-scale organization of metabolic networks, Nature, 407: 651-654.

[6] Enugala, R., Rajamani, L., Ali, K., Kurapati, S. 2015. Community detection in dynamic social networks: A survey. International Journal of Research and Applications, 2(6): 278-285.

[7] Pizzuti, C. 2008. GA-Net: A Genetic Algorithm for Community Detection in Social Networks. PPSN, 51(99): 1081-1090.

[8] Raju Enugala, Lakshmi Rajamani, Sravanthi Kurapati, Mohammad Ali Kadampur, and Rama Devi, Y. 2018. Detecting communities in dynamic social networks using modularity ensembles SOM. International Journal of Rough Sets and Data Analysis (IJRSDA), 5(1): 34-43.

[9] Flake, G. W., Lawrence, S. and Giles, C. L. 2000. Efficient identification of web communities. Proceedings of ACM Conference on Knowledge and Data Discovery (KDD 2000), 150-160.

[10] Newman, M.E.J. and Girvan, M. 2004. Finding and evaluating community structure in networks. Physical Review E 69.

[11] Girvan, M. and Newman, M. E. J. 2002. Community structure in social and biological networks. Proceedings of National Academy of Sciences. USA, 99(12): 7821-7826.

[12] Hopcroft, J.E., Khan, O. and Kulis, B. and Selman, B. 2003. Natural communities in large linked networks. Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD 2003), 541-546.

[13] Raju, E., Ramadevi, Y. and Sravanthi, K. 2018. CILPA: a cohesion index based label propagation algorithm for unveiling communities in complex social networks. International Journal of Information Technology, 10: 435-445.

[14] Raju, E., Rama Devi, Y. and Sravanthi, K. 2017. CCLPA: A clustering coefficient-based label propagation algorithm for unfolding communities in complex networks. 2nd International Conference on Communication and Electronics Systems (ICCES). 240-245.

[15] Niu, S.H., Wang, D., Feng, S.H. and Yu, G. 2009. An improved spectral clustering algorithm for community discovery. Ninth international Conference on Hybrid Intelligent Systems. 262-267.

[16] Raju, E., Hameed, M. A., and Sravanthi, K. 2015. Detecting Communities in Social Networks using Unnormalized Spectral Clustering incorporated with Bisecting K-means. Proceedings of 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2: 903 – 908.

[17] Zachary, W. 1977. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 33: 452-473.

[18] http://vlado.fmfuni-Ij.si/pub/networks/pajek/data/gphs.html.