Network Entropy based Overlapping Community Detection in Social Networks

S Rao Chintalapudi
Department of CSE
University College of Engineering Kakinada (A)
JNTU Kakinada, Kakinada
srao.chintalapudi@gmail.com

M H M Krishna Prasad
Department of CSE
University College of Engineering Kakinada (A)
JNTU Kakinada, Kakinada
krishnaprasad.mhm@jntucek.ac.in

ABSTRACT
The structural analysis of Social networks is gaining more importance over recent years. The most important structural property of social network is community structure and to detect such structures a novel approach is proposed by adopting the information theoretic definition of Entropy to Networks i.e Network Entropy. It is observed that the quality of the community is decreased with higher values of network entropy. Hence, the proposed approach Entropy based Overlapping community detection (EOCD) finds communities that are having low network entropy. EOCD is tested on both real-world and synthetic datasets and the results are evaluated with three popular metrics F-score, Extended NMI and Overlap modularity. It produces high quality communities even for larger networks and delivers good performance over its baseline algorithms.

Keywords
Graph based Data Analysis; Social Network Analysis; Network Science; Community Detection; Network Entropy; Overlapping Communities.

1. INTRODUCTION
Network analysis is an emerging area in which data can be represented in the form of nodes and links. There are several types of networks can be formed in different sectors such as Social Networks, Co-authorship Networks, Transportation Networks, Citation Networks, Communication Networks and so on. The most important property of these networks is its community structure. Community [1] is a collection of more densely connected nodes than the rest of network. Communities are formed naturally as a consequence of simple interactions among people/node and have properties like high internal connections, high modularity, low path length and high robustness. These are very helpful in finding patterns at group level rather than individual level. Nowadays, the problem of detecting communities is crucial for several real time applications like Recommendation systems and Target Marketing.

ICC '17, March 22 2017, Cambridge, United Kingdom

DOI: http://dx.doi.org/10.1145/3018896.3025161

Many researchers proposed several algorithms [1] [2] such as modularity optimization, label propagation, edge betweenness, infomap …etc to detect communities. There are two types of communities, disjoint communities and overlapping communities. Former one does not share a node between communities and the later shares. The primary focus of this paper is to design an Overlapping community detection algorithm using information theory concept i.e entropy.

The rest of the paper is organized as follows. The review of the related work is discussed in Section 2. In Section 3, entropy based overlapping community detection algorithm is discussed. The experimental setup, datasets used in the experiments and the results are discussed in section 4. Finally, the conclusions and future scope of the work is described in Section 5.

2. RELATED WORK
Juan David Cruz et. al [3] proposed a community detection algorithm based on structural closeness and semantic closeness. Here, structural closeness and semantic closeness is measured using modularity and data entropy respectively. This method maximizes the modularity and minimizes the data entropy of a community. Here, entropy minimization can be done using Monte-Carlo approach. The definition of the entropy used in this approach is shown in Eq.(1).

$$H(C) = -\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} s_{ij} \ln s_{ij} + (1 - s_{ij}) \ln(1 - s_{ij}) \quad (1)$$

Where s_{ij} is similarity between i and j, N is the number of nodes in the community.

Lius Argerich [4] proposed a fast algorithm for small community detection in large graphs called Entropy Walker. It can detect overlapping communities under constraints such as minimum and maximum number of members allowed in a community. It is based on simulation of random walks and measures the entropy of each random walk to detect community.

Yong Li et. al [5] proposed a social network community detection method using Shannon entropy. This approach uses entropy to measure network's information and communities can be detected as process of information loss.

Network Entropy is a new information theory based definition proposed by Edward Casey Kenley et. al [6]. Consider an undirected Network N is decomposed into several communities. A community of the network has dense intra connections within the community and sparse inter connections to the rest of the network. In links of a node can be defined as number of connections within the community, whereas out links of a node is the number of

connections outside the community. The probability of a node n having in links can be defined as in Eq. (2).

$$p_{in}(n) = \frac{Neighbors\_of\_node\_n}{Total\_Neighbors\_of\_n} \quad (2)$$

Similarly, the probability of a node n having out links can be defined as in Eq. (3).

$$p_{out}(n) = 1 - p_{in}(n) \quad (3)$$

Entropy of a node H(n)can be defined using pin(n) and pout(n) as in Eq.(4).

$$H(n) = -p_{in}(n)\log p_{in}(n) - p_{out}(n)\log p_{out}(n) \quad (4)$$

Finally, the sum of all nodes entropy gives the entropy of the given network H(N).

$$H(N) = sum(H(n)) \forall n \in N \quad (5)$$

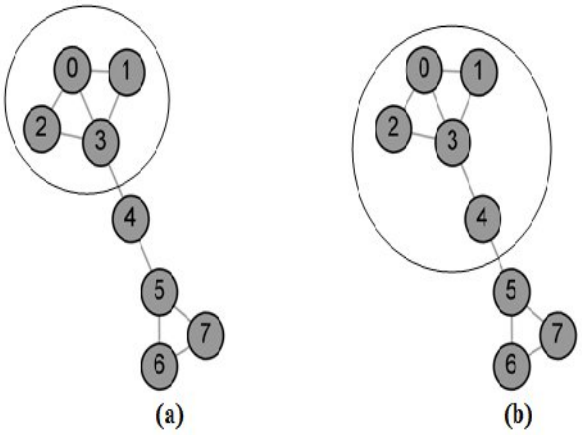

Figure 1. Sample network with 8 nodes (a). nodes {0,1,2,3} are considered as a community. (b). nodes { 0,1,2,3,4} are considered as a community.

Figure 1 (a) (b) illustrates a sample network with 8 nodes. In Figure 1(a), the nodes that are encircled {0,1,2,3} forms a community and its node entropy values are H(0)=H(1)=H(2)=0, H(3)=0.81, H(4)=1.00, H(5)=H(6)=H(7)=0. Hence, the network entropy of Fig .1(a) is 1.81. Similarly, in Fig. 1(b), the nodes that are encircled {0,1,2,3,4} forms a community and its vertex entropy values are H(0)=H(1)=H(2)=H(3)=0, H(4)=1.00, H(5)=0.92, H(6)=H(7)=0.Hence, the graph entropy of Fig .1(b) is 1.92. Here, the observation is that including node id 4 into community increases network entropy from 1.81 to 1.92. That means, whenever the value of network entropy increases the quality of community will be degraded.

3. ENTROPY BASED OVERLAPPING COMMUNITY DETECTION (EOCD)

The problem statement is that for a given Network N, the task of community detection is to find groups of nodes in such a way that the edges within the group are more dense than the between other groups. Here, the overlapping community is that a node can be a member of more than one community.

The major steps involved in this algorithm are given below.

Input **:** Network in Edge list format

Output : Communities

Step 1: Select a node randomly as a seed node.

Step 2: Form an initial community by including selected seed node and its neighbours.

Step 3: Remove nodes iteratively to minimize the network entropy.

Step 4: Add nodes iteratively that are outside the community but having links to the community to minimize network entropy.

Step 5: write the community in the output file.

Step 6: Repeat Step 1-4 until all nodes are placed at least one community.

In the above algorithm, step-1 selects a node from the list of nodes in random manner and in step-2, the neighbours of the seed node will be listed and forms an initial community with seed node and its neighbours. In step-3, each seed neighbour is checked using greedy approach to minimize network entropy with the removal of a node. Similarly, in step-4, the node outside the community but having edges to the community members is added into community if it decreases network entropy. Step-1 to Step-4 detects a locally optimal community with lowest network entropy. Step-5 writes the community in the output file. Repeat step-1 to 5 to generate a set of communities and the nodes in the previously generated communities are retain in the network. That means the same nodes are members of subsequent communities also. In this way, the proposed algorithm detects overlapping communities. There is flexibility in the algorithm that it can detect disjoint communities also. It can be done by removing nodes in each community found at step 5 from the candidate list.

Due to random selection of seed node in the proposed approach, it may produce different output at different runs. To improve the algorithm over its randomness, authors suggested to adopt the concept of cores in the network. There are several measures to find cores such as degree centrality, clustering coefficient. Here, authors adopted degree centrality to find the cores of the network and the nodes can be arranged in descending order of degree centrality. In Seed node selection, priority can be given to nodes with higher degree centrality. The algorithm still suffers with randomness because nodes can be added to the community at random in step-4. This can be solved by adding node with lowest entropy to the community using deterministic greedy approach rather than random manner. This step makes the algorithm non deterministic to deterministic.

4. EXPERIMENTS AND RESULT ANALYSIS

All the experiments are conducted on a Intel Xeon E5-2620 CPU at 2.10 GHz with main memory of 32 GB. The algorithm is implemented in java and the analysis part is done by using R. The results are reviewed by applying EOCD method to Real-world networks as well as Synthetic networks and are discussed in the following sections.

4.1 Tests on Synthetic Networks

Firstly, the algorithm is experimented on Lancichinetti - Fortunato- Radicchi (LFR) benchmark networks [8] seven benchmark networks are generated by varying the value of number of communities a node participates (O_m) from 2 to 8 with the fixed parameters such as number of nodes (N=5000), average

degree($\bar{k}$=10), mixing parameter (μ=0.1), maximum degree (maxk=30), exponents of the power law distribution of vertex degrees and community sizes (t_1=2, t_2=1), minimum community size (minc=10), maximum community size (maxc=50) and number of overlapping nodes (O_n=100). These seven LFR benchmark networks are evaluated using F-score and Extended NMI.

4.1.1 F-score

F-score is used to find the accuracy of the community detection and is evaluated by comparing detected communities and ground truth communities. To measure the accuracy of the proposed algorithm, F-score is used. It is the harmonic mean of Precision and Recall and it can be computed using Eq. (6), Eq. (7) and Eq. (8).

$$precision = \frac{D \cap G}{D} \tag{6}$$

$$recall = \frac{D \cap G}{G} \tag{7}$$

Where D represents detected communities and G represents ground truth communities.

$$F - score = \frac{2 * precision * recall}{precision + recall} \tag{8}$$

The performance of the algorithm is studied in terms of F-score by varying Om and it is compared with two overlapping algorithms SLPA [9] and COPRA [10] as in Fig.2. It depicts that the values of F-score for EOCD are higher than the SLPA and COPRA. And it is also observed that EOCD is stable for even higher values of Om. Hence, EOCD is efficient in detecting overlapping communities and useful for several recommendation systems.

4.1.2 Extended Normalized Mutual Information

Extended Normalized Mutual Information is a variant of normalized mutual information used for evaluating overlapping community detection algorithms proposed by Lancichinetti et.al [11]. The range of values is 0 to 1, where 1 indicates both the covers are identical.

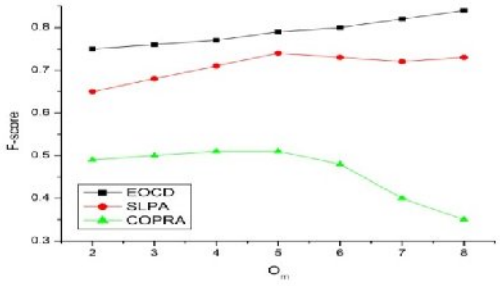

Figure 2. F- Score for LFR Benchmark Network with N=5000, $\bar{k}$ =10, μ =0.1.

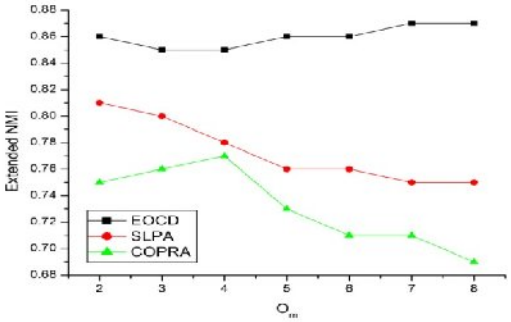

Figure 3. Extended NMI for LFR Benchmark Network with N=5000, $\bar{k}$ =10, μ =0.1.

The quality of the communities detected by EOCD is assessed using extended normalized mutual information and compared with two overlapping algorithms SLPA and COPRA as in Fig. 3. It shows the quality of communities is closed to ground truth and higher values than the other two algorithms.

4.2 Tests on Real-World Networks

Seven Real-world networks are used to test the performance of the algorithm and the summary of these datasets is shown in Table. 1. The first four networks such as Karate [12], Dolphins [13], Football [14], Pollblogs [15] are classic networks that are commonly used for evaluating community detection algorithms and are very small in size. The remaining three networks namely Amazon, DBLP, YouTube are large in size and are acquired from Stanford Large Network Dataset Collection [16] (SNAP Datasets), which contains a collection of publicly available real-world networks. These three networks are mainly used to test the scalability of the algorithm. Ground truth is essential to compute F-Score and NMI but it is typically unavailable for all real-world networks. Hence, evaluation is done using Overlap Modularity.

4.2.1 Overlap Modularity

Overlap Modularity is a measure proposed by Nicosia et.al [17] to evaluate overlapping communities and is an extension of Newman's modularity. It is defined in terms of a function F, which in turn is defined as f(x)=60x-30, a linear scaling function. Overlap modularity is computed for the results obtained from the algorithm on real-world networks and compared with its competing algorithms namely SLPA, COPRA. The values of overlap modularity obtained from three algorithms are listed in Table. 2. It is observed that the proposed algorithm gives higher values of Overlap modularity than the SLPA and COPRA. Hence, the proposed approach produces highly modular structures for real-world networks. And the algorithm is also tested on larger real-world networks such as Amazon, DBLP, YouTube to check the scalability of the algorithm. Finally, the proposed algorithm detects overlap communities even in networks with million number of nodes.

Table 1. Summary of Real-world Network Datasets

Dataset	|V|	|E|
Karate	34	78

Dolphins	62	159
Football	115	613
Pollblogs	1490	16718
Amazon	334863	925872
DBLP	317080	1049866
YouTube	1134890	2987624

Table 2. Overlap Modularity of different algorithms on Real-world networks

Dataset	EOCD	SLPA	COPRA
Karate	0.68	0.65	0.44
Dolphins	0.80	0.76	0.70
Football	0.74	0.70	0.69
Pollblogs	0.79	0.72	0.67
Amazon	0.52	0.46	0.44
DBLP	0.45	0.41	0.34
YouTube	0.48	0.43	0.40

5. CONCLUSION AND FUTURE SCOPE

In this paper, a new definition of entropy is adopted to detect overlapping communities in social networks. The proposed algorithm Entropy based Overlapping Community Detection (EOCD) is tested on both LFR benchmark networks and real-world networks. The performance of the algorithm is evaluated in terms of F-score, Extended NMI and Overlap Modularity. The scalability of the algorithm also tested using larger real-world networks like Amazon, DBLP, Youtube. EOCD uncovers the latent community structures effectively from complex networks and the quality of the detected communities is good over its peer algorithms SLPA and COPRA. It has many potential application areas such as social network analysis, biology, target marketing and recommendation systems. The execution time for networks with millions of nodes is high because of network entropy computation. Hence, it can be further improved using High Performance Computing with GPUs to reduce the execution time even for larger networks in future.

6. REFERENCES

[1] Harenberg, S., Bello, G., Gjeltema L., Ranshous, S., Jitendra, H., Seay, R., Padmanabhan, K., and Samatova N. 2014. Community detection in large-scale networks: a survey and empirical evaluation. *WIREs Comput Stat*. 6, 6 (Jul. 2014), 426-439. DOI= http://dx.doi.org/10.1002/wics.1319.

[2] Rao Chintalapudi S., and Krishna Prasad M. H. M. 2015. A survey on community detection algorithms in large scale real world networks. *In Proceedings of 2nd International conference on computing for sustainable global development*. 1323-1327.

[3] Cruz J. D., Bothorel C., and Poulet F. 2011. Entropy Based Community Detection in Augmented Social Networks. *In proceedings of International Conference on Computational Aspects of Social Networks*.163-168. DOI= 10.1109/CASON.2011.6085937.

[4] Argerich L. 2015. Entropy Walker, a fast algorithm for small community detection in large graphs. *CoRR* 1505.02406.

[5] Li Y., Zhang G., Feng Y., and Wu C. 2014. An entropy-based social network community detecting method and its application to scientometrics. *Scientometrics*.102, 1 (Jul. 2014), 1003-1017. DOI= 10.1007/s11192-014-1377-5.

[6] Kenley E. C., Cho Y. 2011. Entropy-Based Graph Clustering: Application to Biological and Social Networks. *In proceedings of 11th IEEE International Conference on Data Mining*.1116-1121. DOI= 10.1109/ICDM.2011.64.

[7] R Development Core Team. 2014. R: A Language and Environment for Statistical Computing. http://www.R-project.org. Accessed 22 June 2014.

[8] Lancichinetti A., Fortunato S., and Radicchi F. 2008. Benchmark graphs for testing community detection algorithms. *Phys Rev E*. 78, 4 (Oct. 2008), 46110. DOI= 10.1103/PhysRevE.78.046110.

[9] Xie J., Szymanski B. K., and Liu X. 2011. SLPA: Un covering Overlapping communities in Social Networks via A Speaker- Listener Interaction Dynamic Process. *In proceedings of 11th IEEE International Conference on Data Mining Workshops*. 344-349. DOI= 10.1109/ICDMW.2011.154.

[10] Gregory S. 2010. Finding overlapping communities in networks by label propagation. *New J Phys*. 12, 10 (Oct. 2010), 103018.

[11] Lancichinetti A., Fortunato S., and Kertész J. 2009. Detecting the overlapping and hierarchical community structure of complex networks. *New J Phys*. 11, 3 (Mar. 2009), 033015. DOI= 10.1088/1367-2630/11/3/033015.

[12] Zachary W. W. 1977. An information flow model for conflict and fission in small groups. *J Anthropol Res*. 33, 452-473.

[13] Lusseau D., Schneider K., Boisseau O. J., Haase P, Slooten E., and Dawson S. M. 2003. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav Ecol Sociobiol*. 54, 396-405.

[14] Girvan M., Newman M. E. J. 2002. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99, 12 (Apr. 2002), 7821-7826.

[15] Adamic L. A., Glance N. 2005. The political blogosphere and the 2004 US Election. *In proceedings of 3rd International Workshop on Link Discovery*. 36-43.

[16] Leskovec J., and Krevl A., SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data. Accessed 15 May 2015.

[17] Nicosia V., Mangioni G., Carciolo V., and Malgeri M. 2009. Extending the definition of Modularity to directed graphs with overlapping communities. *J Stat Mech-Theory E*. 2009, 3 (Mar. 2009), P03024. DOI= 10.1088/1742-5468/2009/03/P03024.