ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Multi Comm_Plus: A Community Detection System for Identification of Community in Multi-Dimensional Networks

Dhanya Sudhakaran[1], Shini Renjith[2]

PG Scholar, Dept. of CSE, Sree Buddha College of Engineering, Alappuzha, India[1]

Assistant Professor, Dept. of CSE, Sree Buddha College of Engineering, Alappuzha, India[2]

ABSTRACT: Community detection is a common problem in graph and big data analytics. It consists of finding groups of densely connected nodes with few connections to nodes outside of the group. In particular, identifying communities in large-scale networks is an important task in many scientific domains. Community detection algorithms in literature proves to be less efficient, as it leads to generation of communities with noisy interactions. To address this limitation, there is a need to develop a system which identifies the best community among multi-dimensional networks based on relevant selection criteria and dimensionality of entities, thereby eliminating the noisy interactions in a real-time environment**.**

KEYWORDS: Community Detection; MultiComm; Tensor; Affinity; Local Modularity Measure; Context- aware Relation.

I. INTRODUCTION

Detecting clusters or communities in large real-world graphs such as large social or information networks is a problem of considerable interest. Community detection is a common area in graph data computations and data mining computations [1] [2]. It consists of finding groups of densely connected nodes with few connections to nodes outside of the group. Networks can be either multi-dimensional networks or uni-dimensional networks. Multi-dimensional networks are networks with multiple kinds of relations. Examples of multi-dimensional networks are social networks, genetic networks, co-citation networks. Each node in a network is an item corresponding to a dimension or entity in a network and each edge indicates a relationship between two nodes. In social networks, finding a community structure means finding a group of users who interact on different entities like tags, photos, comments or stories. In case of a co-citation network, community structure represents a group of authors who interact on publication information such as titles, abstracts, keywords etc. Detecting communities is of great importance in sociology, biology, computer science etc. In particular, identifying communities in large-scale networks is an important task in many scientific domains. Large-scale networks with thousands to millions of nodes are common across many scientific domains. Finding community structures from these networks are of particular interest. Identifying communities in a large-scale network is a complex task because there exists many definitions of community and intractability of the community detection algorithms. Community detection in multi-dimensional networks is restricted to the dimensionality of the entity structure, selection criteria and the current requirements of the entities are to be considered. There is a need to recover the community from noisy interactions among the entities. The community detection problem has many widespread applications and has hence proven to be very important. This paper is organized as follows. Section II presents some of the related works in the area of community detection. Section III presents the existing system and architecture. Section IV presents the proposed system. Concluding remarks are given in Section IV

II. RELATED WORK

A community could be loosely described as a collection of vertices within a graph that are densely connected amongst them while being loosely connected to the rest of the graph. Bagrow and Bolt [7] proposed a local method for

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

detecting communities. They proposed an algorithm which consists of a shell 'l' spreading outward from a starting vertex. As the starting vertex's nearest neighbors and next nearest neighbors etc. are visited by the shell 'l', two quantities are computed: emerging degree and total emerging degree. Algorithm works by expanding the shell outward from some starting vertex 'j' and comparing the total emerging degree change to some threshold value. When the 'l' shell ceases to grow, all vertices covered by shells of a depth <= l are listed as members of vertex j's community. Ruan and Zhang [8] proposed a quantitative measure called modularity to assess the quality of community structures. Modularity means the measure of fraction of edges falling within communities subtracted by what one would expect if the edges are randomly placed. It provides a good quality measure to compare different community structures. A larger modularity value means stronger community structures. Newman [9], Duch and Arenas [10] proposed an algorithm by optimizing the modularity measure.

A graph partition method based on min-max clustering principle was proposed by Ding and Zha et al. [11]. The principle states that the similarity or association between two subgraphs is minimized, while the similarity or association within each subgraph is maximized. Luo and Wang et al. [12] proposed a framework to identify modules within a biological network. Networks are divided into sub-networks and the identification of modules is based on their topology. For this, the concept of edge-betweenness was used. Edge-betweenness is the number of shortest path between all pairs of vertices that run through the edge. Edges between modules tend to have shortest paths through them than do edges inside modules and thus have higher betweenness values. The deletion of edges with high betweenness can separate the network, while keeping the modules structure in the network intact. Sun and Castro et al. [13] proposed a framework, MetaFac that extracts community structures from social media networks. Mehler and Skicna [14] presented a general method for network community expansion from seed set of members. It is achieved by assigning a score to all entities in the network and selecting the highest scoring outside vertex to join the community. Some of the scoring criteria in order to rank the selection are neighbor count, juxta position count, neighbor ratio, juxta position ratio, binomial probability. The essential function of the community expansion method is to identify the most promising next member to be added to the community. Some representative community detection methods [15] such as latent space models, block model approximation, spectral clustering and modularity maximization.

Adaptive algorithms were developed for detecting community structures in dynamic social networks. Quick Community Adaptation (QCA) [16] is an adaptive modularity based method for identifying and tracing community structure in dynamic social networks. Modularity based approaches are used for finding community structures in very large networks. Modularity is a property of a network and a specified proposed division of that network into communities. Clauset and Newman [17] proposed an algorithm based on the modularity property of the network. Chikhi and Rothenburger [18] proposed a probabilistic approach known as Smoothed Probabilistic Community Explorer (SPCE), a generative model for community structure identification. SPCE provides several advantages. It finds coherent and overlapping community structures. It takes as input only the number of communities to identify and not their size. It detects communities in directed and undirected networks. It provides a two-view community structure in directed networks and is able to analyze weighted and unweighted networks. Usha and Reka et al. [19] proposed a localized community detection algorithm based on label propagation. For finding overlapping communities in large networks label propagation method [20] can be used. Chang and Yi-Hsu et al. [21] also developed a general probabilistic framework for detecting community structure. Key idea of generalization is to characterize a network by a bivariate distribution that specifies the probability of the two vertices appearing at both ends of a randomly selected path in the graph. Riedy and Bader et al. [22] presented a greedy agglomerative algorithm that grows a community around a given small seed set. Starting from a set of seed vertices the algorithm pull adjacent vertices into the community to maximize modularity. Random walk process can be used to compute communities in large networks. Such an algorithm known as walktrap was proposed by Pons and Latapy [23]. Improved community detection algorithm based on random walk by taking into account node attribute information was proposed by Daxiang and Sun et al. [24] Random walk process can also be used for detecting community structures for undirected graphs [25]. For finding overlapping communities Ball and Karrer et al. [26] proposed a method based on a statistical approach using generative network models. Algorithms named RaRe (Rank Removal) and IS (Iterative Scan) were proposed by Baumes and Goldberg et al. in [27]. IS iteratively constructs clusters and RaRe attempts to identify high ranking nodes and remove them from graph, in order to disconnect the graph into smaller connected components.

A visual data mining approach to find overlapping communities in networks was proposed by Chen and Osman et al. The proposed algorithm was known as ONDOCS (Ordering Nodes to Detect Overlapping Community Structure), helps the user to make appropriate parameter selections by observing initial data visualizations and finds and extracts overlapping community structures from the network. A game theoretic framework to address the community detection

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

based on the structures of the social networks was proposed [28] to find the overlapping communities. Community formation s formulated as a strategic game known as community formation game. A modification to Cluster Overlap Newman- Girman Algorithm (CONGA) was proposed by Gregory [29] known as CONGA Optimized (CONGO). Overlapping communities were detected by using the concept of split-betweenness. A two-phase method of overlapping communities was proposed in [30] known as Peacock Algorithm. In the first phase, a network is transformed to a new one by splitting vertices using the idea of split betweenness. In the second phase, the transformed network is processed by a disjoint community detection algorithm. This approach had the potential to convert any disjoint community detection algorithm into an overlapping community detection algorithm.

III. MULTICOMM

In [31], proposed a framework known as MultiComm, for finding the community structure in multi-dimensional networks. Multi-dimensional networks are networks with multiple kinds of relations. Examples of such networks include social media networks, co-citation networks, genetic networks etc. In case of the online social networks, finding a community structure means finding a group of users who interact on different data entities. These entities or dimensions can be users, tags, photos, comments etc. In case of a co-citation network, finding a community structure means to find a group of authors who relate to other authors who interact significantly on publication information such as titles, abstracts, keywords. The main aim was to propose an algorithm, MultiComm, to identify a seed-based community structure in a multi-dimensional network such that the involved items of the entities inside the community interact significantly, and meanwhile, they are not strongly influenced by the items outside the community. In the proposal, a community is constructed starting with a seed consisting of one or more items of the entities believed to be participating in a viable community.

Given the seed item, iteratively adjoin new items by evaluating the affinity between the items to build a community in the network. As there are multiple interactions among the items from different dimensions/entities in a multidimensional network, the main challenge is how to evaluate the affinity between the two items in the same type of entity (from the same dimension/entity) or in different types of entities (from different dimensions/entities). For example, consider an academic publication network, where some concepts are labelled to papers and paper is associated with several keywords and authors. So there are four dimensions or entities in this network such as, paper, concept, author, and keyword. Items in three dimensions i.e., author, keyword and paper are related among themselves and the items in two dimensions i.e., paper and concept are related to each other. A tensor can be used to represent the interactions among these entities. In Figure 1, the affinity between a paper “A” and a paper “B” (the same type of entity), and the affinity between a paper “A” and a keyword “C” (different types of entities) are required in order to evaluate and decide the papers “A” and “B” or the paper “A” and the keyword “C” to put together in a community. On the other hand, a criterion is needed in order to evaluate a high quality of generated communities by the proposed algorithm, and by a local modularity measure of a community in a multi-dimensional network.

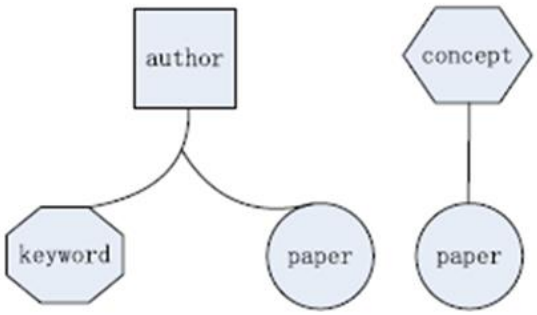

Figure 1: Interaction between entities

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

After solving a set of tensors, obtain the probability distributions of visiting each item in each dimension in the multi-dimensional network. These probability distributions can be viewed as an affinity vector because it indicates the affinity of the items in each dimension to the items in the current community. Based on their probability values, the candidate items in different dimensions that are closely related to the current items in the community can be determined. This affinity vector is also known as transition probability tensor. In order to determine the best community, a local modularity measure can be used.

A. *MultiComm Architecture:*

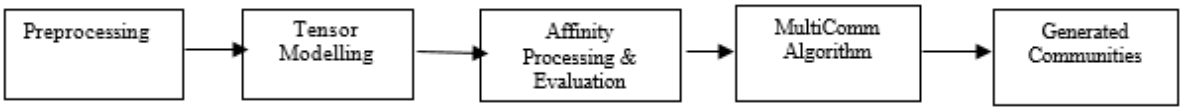

Figure 2: MultiComm System Architecture

The first step is to preprocess the data and then forming a set of tensors. Next step is the affinity processing and evaluating the affinities of the communities. Next step is the implementation of the MultiComm Algorithm to generate the resultant communities.

B. *MultiComm Algorithm:*

Input: transition probability tensors and a set of seed items (L) and its related dimensions.
Output: A set of items for a community.
Procedure
1. Obtain the transition probability vectors xv.
2. For each dimension, find the candidate item which is not in L and has the highest probabilities in transition probability vectors and generate the set of candidate items from all dimensions to Lc.
3. For each candidate item in Lc, determine the change of local modularity measure of its corresponding tensor after this candidate item is added.
4. If the increases of local modularity for all the candidate items in Lc are not significant, then stop. Otherwise select the item which has the largest increase of modularity and go to step 1.

IV. PROPOSED SYSTEM

A. *Methodology*

1. First step is to preprocess the data. Data is the author documents which are considered as multi-dimensional (paper, author, keyword, paper category, concepts etc.). Consider an academic publication network where some concepts are labeled to papers and each paper is associated with several keywords and authors. So there are four dimensions or entities in this network i.e., paper, concept, author and keyword. Items in three dimensions i.e., author, keyword and paper are related among themselves and items in two dimensions i.e., paper and concept are related to each other.
2. Next step is tensor modeling and affinity processing. Tensors are used to represent interactions among the entities in the multi-dimensional networks. For example, author-paper-keyword tensor, paper-keyword tensor etc. A tensor can be used to represent the interactions among the entities.
3. During the affinity processing, obtain the probability distribution of visiting each item in each dimension and it can be viewed as an affinity vector, indicating the affinity of each items in the current community.
4. Based on the probability values, candidate items in different dimensions that are closely related to the current items in the community can be determined.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

5. Next step is the implementation of MultiComm Algorithm.
6. In MultiComm, community is constructed starting with a seed consisting of one or more entities given by the user. Given the seed items, iteratively adjoin new items by evaluating the affinity between the items to build a community in the network.
7. A context- aware relation extraction method can be used for Relation Completion task. CORE employs Relation Query (RelQuery) which is basically a web search query that is specially formulated for the purpose of relation completion

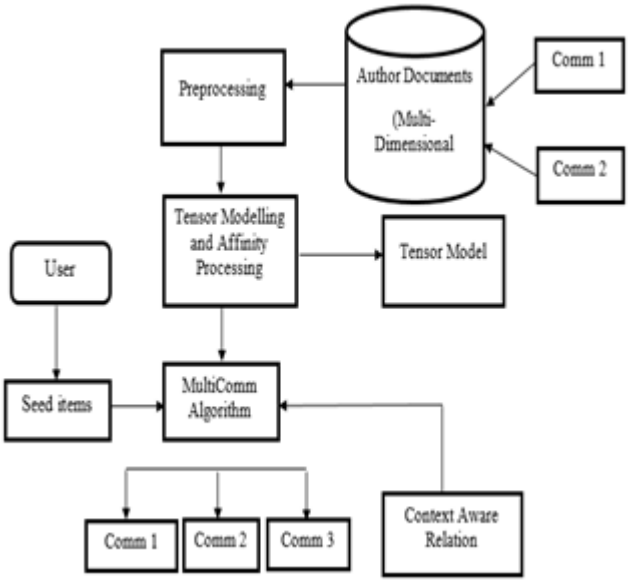

Figure 3. Proposed system architecture

V. RESULT

Difference between the system design of MultiComm and MultiComm_Plus can be identified from the figure 1 and figure 3. The proposed system architecture involves a context aware relation scheme which is not present in the MultiComm.

VI. CONCLUSION

Community detection algorithms are widely used to study the structural and topological properties of real-world networks. Identification of the best community among the network based on the current scenario is a big challenge. Current system generates communities with some noisy interactions. So there is need to develop a system to overcome this limitation. MultiComm Algorithm can be enhanced with a context-aware relation extraction method, so that generated communities can be recovered from noisy interactions.

REFERENCES

1. ShiniRenjith, C Anjali, "Fitness function in genetic algorithm based information retrieval: A Survey", ICMIC13, December 2013, pp. 80-86.
2. DhanyaSudhakaran, ShiniRenjith, "Phase based resource aware scheduler with job profiling for MapReduce", IJLTET, vol 6, Issue 2, Nov 2015, pp.92-96.
3. F Moradi, T Olovsson, P Tsigas, "An of community detection algorithms on large-scale email traffic", in: SEA. Berlin/Heidelberg: Springer; 2012; 283–294.
4. J Leskovec, K.J Lang, M.W Mahoney," Empirical comparison of algorithms for network community detection", CoRR, abs/1004.3539, 2010.
5. A Lancichinetti, S Fortunato, "Community detection algorithms: a comparative analysis", Phys Rev E 2009, 80:056117.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

6. F.D Malliaros, M Vazirgiannis, "Clustering and community detection in directed networks: a survey", CoRR, abs/1308.0971, 2013.
7. J. Bagrow and E. Bolt, "Local Method for Detecting Communities," Physical Rev. E, vol. 72, no. 4, p. 046108, 2005.
8. J. Ruan and W. Zhang, "An Efficient Spectral Algorithm for Network Community Discovery and Its Applications to Biological and Social Networks," Proc. Seventh IEEE Int'l Conf. Data Mining (ICDM '07), pp. 643-648, Jan. 2007.
9. M. Newman, "The Structure and Function of Complex Networks," SIAM Rev., vol. 45, no. 2, pp. 167-256, 2003.
10. Jordi Duch, Alex Arenas, "Community detection in complex networks using extremal optimization", Phys, Rev E72, 027104, August 2005.
11. C.H.Q. Ding, X. He, H. Zha, and M. Gu and H.D. Simon, "A Min- Max Cut Algorithm for Graph Partitioning and Data Clustering," Proc. IEEE Int'l Conf. Data Mining, pp. 107-114, 2001.
12. F. Luo, J.Z. Wang, and E. Promislow, "Exploring Local Community Structures in Large Networks," Web Intelligence and Agent Systems, vol. 6, no. 4, pp. 387-400, 2008.
13. Y. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram, and A. Kelliher, "Metafac: Community Discovery via Relational Hypergraph Factorization," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '09), pp. 527-536, 2009.
    a. Mehler and S. Skiena, "Expanding Network Communities from Representative Examples," ACM Trans. Knowledge Discovery from Data, vol. 3, no. 2, article 7, 2009.
14. Lei Tang, Xufei Wang, Huan Liu, "Community Detection in Multi-Dimensional Networks", Technical Report, TR-10-006, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287, 2010.
15. Nam P. Nguyen, Thang N. Dinh, Ying Xuan, My T. Thai, "Adaptive Algorithms for Detecting Community Structure in Dynamic Social Networks", IEEE infocom, 2011.
16. Aaron Clauset, M. E. J. Newman, and Cristopher Moore, "Finding community structure in very large networks", Phy Rev E70, vol.6, December 2008.
17. NacimFatehChikhi, Bernard Rothenburger, Nathalie Aussenac-Gilles, "Community Structure Identification: A Probabilistic Approach", conference paper, IEEE explore January 2010.
18. UshaNandiniRaghavan, R eka Albert, and Soundar Kumara, "Near linear time algorithm to detect community structures in large-scale networks", Physical Review E, 76(3):036106, 2007.
19. Steve Gregory, "Finding overlapping communities in networks by label propagation", New journal of Physics, October 2010.
20. Cheng-Shang Chang, Chin-Yi Hsu, Jay Cheng, and Duan-Shin Lee, "A General Probabilistic Framework for Detecting Community Structure in Networks"
21. Jason Riedy, A. David Bader Karl, Jiang PushkarPande, Richa Sharma," Detecting Communities from Given Seeds in Social Networks", February 22, 2011.
22. Pascal Pons and MatthieuLatapy, "Computing communities in large networks using random walks", ISCIS, Springer, vol 3733, 2005, pp.284-293.
23. DaxiangJi, Yuqing Sun and Demin Li, " Improved Random Walk Based Community Detection Algorithm", International Journal of Multimedia and Ubiquitous Engineering Vol.9, No.5 2014, pp.131-142.
24. Xiaoming Liu, Yadong Zhou, Chengchen Hu, Xiaohong Guan, JunyuanLeng, "Detecting Community Structure for Undirected Big Graphs Based on Random Walks", www'14 Proceedings of 23rd international conference on WWW, 2014, pp.1151-1156.
25. Brian Ball, Brian Karrer, M.E.J Newman, "An efficient and principled method for detecting communities in networks", April 2011.
26. Jeffrey Baumes, Mark Goldberg, MukkaiKrishnamoorthy, "Finding communities by clustering a graph into overlapping subgraphs," IADIS International Conference on Applied Computing 2005.
27. Wei Chen, Zhenming Liu, Xiaorui Sun, Yajun Wang, "A game-theoretic framework to identify overlapping communities in social networks", Data Min Knowl Disc (2010) 21:224–240.
28. S. Gregory, "A fast algorithm to find overlapping communities in networks," in PKDD, 2008, pp. 408–423.
29. S. Gregory, "An algorithm to find overlapping community structure in networks," in PKDD, 2007, pp. 91–102.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
30. Xutao Li, Michael K. Ng, and Yunming Ye, "MultiComm: Finding Community Structure in Multi-Dimensional Networks", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 4, April 2014.

BIOGRAPHY

DhanyaSudhakaran is an M. Tech Post Graduate Scholar specialized in Computer Science and Engineering. She has pursued B.Tech in Computer Science and Engineering in 2013 under the University of Kerala. She has published a paper in the area of data mining and MapReduce. Her areas of interest include data mining, Big data Analytics, Cryptographic Security. Her research area is community detection in multi-dimensional networks. She is a member of Association for Computing Machinery (ACM) and Computer Society of India (CSI).

ShiniRenjith has M.Tech in Computer Science from University of Kerala (2014) and B.Tech in Computer Science and Engineering from Cochin University of Science and Technology, CUSAT (2004). Also she is currently pursuing her PhD from CUSAT and is working as an Assistant Professor at Computer Science and Engineering Department in Sree Buddha College of Engineering, Alappuzha. Prior to this she has worked at College of Engineering, Munnar and College of Engineering, Thiruvananthapuram. She is qualified in UGC NET, has won the Best Paper Award at International Conference on Mobility in Computing [ICMiC13] and has presented and published 6 papers in various international conferences and journals. Her current areas of interests include big data analysis, Information filtering and computer networks. She is an active member of Association for Computing Machinery (ACM) and IEEE.