A Community Detection Algorithm Based on Granulation of Links

Samrat Gupta
Information Systems Department
Indian Institute of Management
Ahmedabad, India
e-mail: samratg@iima.ac.in

Pradeep Kumar
IT and Systems Department
Indian Institute of Management
Lucknow, India
e-mail: pradeepkumar@iiml.ac.in

Irina Perfilieva
Centre of Excellence IT4Innovations
University of Ostrava
Ostrava, Czech Republic
email: irina.perfilieva@osu.cz

***Abstract*—The digital transformation of business and society has led to the growth of networks in almost every field. Finding communities in real world networks has been considered crucial for modern network science. Moreover, the organization of communities into co-occurring disjoint, nested and overlapping structures adds to the complexity of community detection problem. Therefore, methodological rigor is crucial for community detection so as to foster cumulative tradition in data and knowledge engineering. This paper proposes an algorithm for overlapping community detection based on the concepts of rough set theory. Initially, subsets of links are formed by using neighborhood links around each pair of nodes. Subsequently, we iteratively obtain the constrained linkage upper approximation of these subsets. The notion of mutual link reciprocity is used as a merging criterion during the iterations. The proposed algorithm is experimentally evaluated on eight real-world networks. Comparative analysis with state-of-the-art algorithms demonstrates the effectiveness of proposed algorithm.**

Keywords- community structure; clustering; rough sets; complex networks

I. INTRODUCTION

The technological advancement in contemporary digital world has led to the formation and mapping of systems which consist of many interconnected dynamical units. Such systems are collectively referred to as complex systems because their constituent units are capable of interacting not only with each other but also with the environment [1]. Some examples of complex systems include a social club that requires cooperation among its members to achieve a common goal, the Internet consisting of millions of interconnected routers and the human brain comprising millions of synaptically connected neurons [2]. All the networks arising from complex systems, irrespective of diversity in their origin, nature, size and scope, follow a common set of organizing principles [1]. Since the existence of cohesive subgroups (generally known as communities) is one of the fundamental properties of complex networks, identifying them is essential to explore and understand the dynamics of complex real-world systems.

Communities are considered as thickly connected subgroups of nodes within a complex network such that the link density within subgroups is much higher than the density of links between subgroups [3]. The dense inter-community connectedness exists due to organizational or functional components within a network such as groups of friends in a social network of students, and groups of companies with interlocking directorates in organizational networks [2]. In this research work, we design a community detection algorithm based on the unexplored theoretical synergy of the concepts of link communities in network science and upper approximation in rough set paradigm. As discussed in Section III, this synergy is constituted by expanding each link and its link neighborhood component using the concept of upper approximation rather than expanding each node and its node neighborhood component which has been the focus in prior research [4]-[7].

This paper is organized as follows. In Section II, we briefly present the motivation behind this work. In Section III, the methodology of the proposed algorithm is explained. Section IV presents the experimental setup. In Section V, the experimental results are discussed. Finally, Section VI concludes this work.

II. MOTIVATION

An effective community detection technique can transform business and society through its widespread applications. These applications range from topic detection in collaborative tagging systems, to event detection on social media content [3]. In the past, community detection techniques have been used to devise antiterrorism strategies and understand functional patterns of the human brain that can help in positioning and pricing of products [8][9].

Though researchers have been addressing the community detection problem for more than a decade, state-of-the-art algorithms still have several limitations [10]. A majority of the existing algorithms assign each node to only one community; some algorithms are domain specific; some require a priori knowledge about the number of communities; some are not scalable for large networks and some are unsusceptible to variations in size of communities. One of the major challenges encountered currently for community detection is the identification of overlapping communities which manifest the reality of today’s world [11]. For instance, in social networks, individuals may belong to multiple communities due to their friendships, professional associations, family relationships and so on. Moreover, communities may overlap not only partially but also entirely such that one community is contained in another [12]. An effective community detection algorithm must be receptive to distinctive features of community structure in complex networks and be capable of detecting co-occurring disjoint, nested and overlapping communities.

ISBN: 978-1-61208-700-9

III. LUAMCOM – PROPOSED ALGORITHM

Soft computing techniques for community detection such as evolutionary computing, swarm intelligence, fuzzy logic and genetic algorithms have gained popularity in recent years [13]-[15]. However, rough set theory has not been explored to its fullest potential for mathematical modelling of complex networks thus indicating methodological gaps in the burgeoning community detection literature.

The proposed Link Upper Approximation Method for COMmunity detection is abbreviated as LUAMCOM. Since, links are more idiosyncratic in nature as compared to nodes, pervasive overlaps in community structure are better discovered by clustering of links rather than nodes [16]. The proposed algorithm considers links of a complex network as entities to be clustered wherein each link and its neighborhood links form initial components (granule in rough set terminology) [17]. These initial components are termed as Link Neighborhood Subsets (LNS). Then, the concept of upper approximation is used iteratively to grow these components in a constrained manner until they merge to form stable components. This step consists of iterative formation of First Linkage Upper Approximation (FLUA) and Constrained Linkage Upper Approximation (CLUA) until convergence. The converged or stable components are actually the identified communities within a network. The notion of Mutual Link Reciprocity (MLR) is used as a merging criterion during the iterations. As shown in Figure 1, the iterative process followed by a fine-tuning process, ensures that the detected community structure displays high intra-community density of links, and low inter-community density of links. To the best of our knowledge, LUAMCOM is the first community detection algorithm based on integration of link granulation and upper approximation.

IV. EXPERIMENTAL SETUP

We conducted experiments on network datasets representing complex systems in diverse domains. These networks consist of a network of friendships at a karate club (34 nodes, 78 links) [18], network of bones in human skull (35 nodes, 79 links) [19], network of associations between dolphins (62 nodes, 159 links) [20], network of friendships in a high school (69 nodes, 220 links) [21], network of political books sold online by Amazon.com (105 nodes, 441 links) [22], co-appearance network of football teams (115 nodes, 613 links) [23], an online network of friendships on Facebook (2888 nodes, 2981 links) and a human protein-protein interaction network (3724 nodes, 8748 links) [24]. All the experimental work was conducted in the R programming environment. We used a system with Intel Core i5 processor and 8.00 GB RAM, running R version 3.2.5. The version 0.8.2 beta of Gephi software has been used for visualization.

To demonstrate the effectiveness of the proposed algorithm, we compare its performance with state-of-the-art community detection techniques using evaluation criteria namely Normalized Mutual Information (NMI) [25], partition density [16] and overlapping modularity [26]. Each of these evaluation measures has its own relevance and importance for measuring the quality of community detection. NMI computes the agreement between detected community structure and true community structure. In this work, we have used a variant of NMI that is designed especially for overlapping communities and results in the same NMI values as the standard measure when there is no overlap. Another measure used in this work is an extended version of modularity that can be used for evaluation of overlapping communities. This measure was proposed to address the limitations of traditional modularity. Finally, we have used a measure called partition to evaluate the quality of link based communities. Since partition density measures the intra-community density of links, it is considered more suitable for assessing the quality of overlapping community detection.

For comparison of the proposed algorithm with state-of-the-art algorithms, we consider the most relevant overlapping community detection algorithms such as Ahn Bagrow Lehmann (ABL) [16], Community Overlap PRopagation Algorithm (COPRA) [27], Clique Percolation Method (CPM) [28] and Greedy Clique Expansion (GCE) [29].

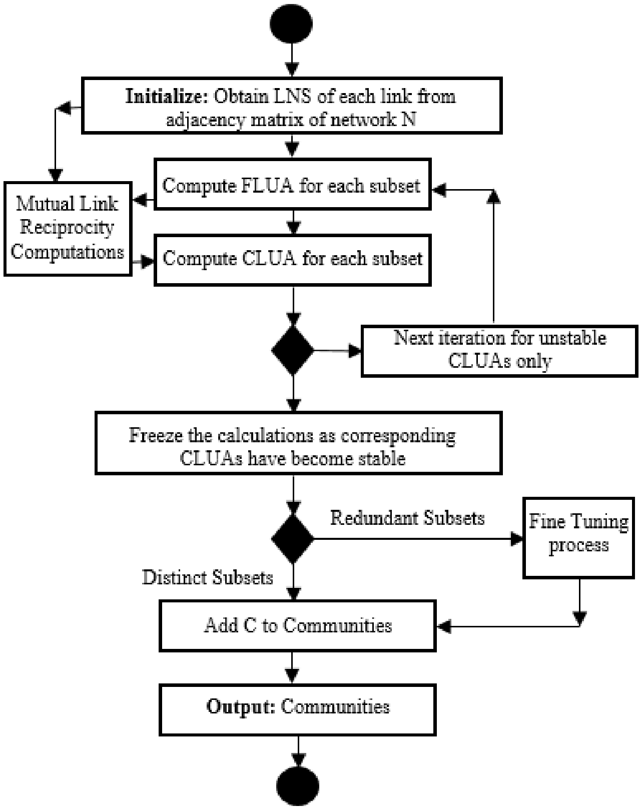
Initialize: Obtain LNS of each link from adjacency matrix of network N
Compute FLUA for each subset
Mutual Link Reciprocity Computations
Compute CLUA for each subset
Next iteration for unstable CLUAs only
Freeze the calculations as corresponding CLUAs have become stable
Redundant Subsets
Fine Tuning process
Distinct Subsets
Add C to Communities
Output: Communities

Figure 1. Flow Diagram of LUAMCOM

V. EXPERIMENTAL RESULTS

The proposed algorithm identifies five overlapping communities in the human skull network. These five communities, generally known as *complexes* in case of biological networks include a complex representing the group of facial bones, groups of cranial bones, cervical bones, left and right ear ossicles bones. The overlapping community structure detected by the proposed algorithm in the human skull network is highly consistent with the results reported in existing literature [19] and shows the functional and developmental dependencies of bones in the human skull.

The proposed algorithm identifies five overlapping communities, one nested and one disjoint community in the high school network. The nested community within grade-9 corresponds to a subgroup of different ethnicity [21]. However, two nodes have been misclassified in the high school network. The proposed algorithm competes favorably with state-of-the-art community detection algorithms. Although the NMI value of GCE is slightly higher than that of the proposed algorithm, it detects only six communities and overlooks the nested community within grade-9 [21]. The overlapping nodes detected by the proposed algorithm are quite similar to existing overlapping community detection algorithms such as COPRA, GCE and ABL.

The experiment on karate club reveals two communities wherein all the nodes have been classified correctly. However, the proposed algorithm detects four nodes as overlapping, thus indicating evidence for dual memberships of some of the members within the karate club. This result is in accordance with the observation that some bridge nodes act as information carriers within the two communities of karate club [18]. The experiment on dolphin network divides the network into two communities consisting of 21 and 41 nodes without any misclassification. However, three overlapping sub-communities were found in the larger community. The composite performance of the proposed algorithm outperforms ABL, CPM and GCE on karate and dolphin networks. However, COPRA obtained higher score on the karate club network because of its higher NMI.

The proposed algorithm identifies three communities with one overlapping node on a network of political books (polbooks). The composite score of the proposed algorithm in the case of the polbooks network is higher than ABL, COPRA, CPM and almost equal to GCE. On the football network, the proposed algorithm identifies 11 communities and outperforms ABL, COPRA and CPM, while being slightly behind GCE in terms of composite score.

We also performed experiments on larger networks such as, a network of friendships extracted from Facebook and human protein interaction network downloaded from an online database [30]. In an undirected version Facebook network, 16 communities identified by the proposed algorithm were found to be structurally similar to those identified by the edge label propagation approach [24]. On the Protein Protein Interaction (PPI) network, the detected community structure was very similar to that of an algorithm based on binary tree theory [31].

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed an algorithm based on granulation and upper approximation of links for overlapping community detection in complex networks. We have investigated the performance of the proposed algorithm on different types of networks. The proposed algorithm performs competitively with state-of-the-art community detection algorithms and effectively detects co-occurring disjoint, nested and overlapping community structures in complex real-world networks. While the proposed algorithm makes an important methodological contribution to the challenging problem of community detection, it is applicable only to undirected and unweighted networks. In the future, we intend to enhance the capability of the proposed methodology to detect communities in directed and weighted networks. We believe that the proposed algorithm will impart rigor while guiding future research in the field of complex network analysis.

REFERENCES

[1] L. A. N. Amaral and J. M. Ottino, “Complex networks,” The European Physical Journal B - Condensed Matter, vol. 38(2), 2004, pp. 147–162. doi: 10.1140/epjb/e2004-00110-5

[2] R. Albert and A.-L. Barabási, “Statistical mechanics of complex networks,” Reviews of Modern Physics, vol. 74(1), 2002, pp. 47. doi: 10.1103/RevModPhys.74.47

[3] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, “Community detection in Social Media: Performance and application considerations,” Data Mining and Knowledge Discovery, vol. 24(3), 2012, pp. 515–554. doi :10.1007/s10618-011-0224-z

[4] P. Kumar, S. Gupta, and B. Bhasker, “An upper approximation based community detection algorithm for complex networks,” Decision Support Systems, vol. 96, 2017, pp. 103-118. doi: 10.1016/j.dss.2017.02.010

[5] Z. Cui, W. Chu, and Y. Fu, “Community structure Detection Algorithm Based on Rough Set,” In Second International Conference on Business Computing and Global Informatization (BCGIN), IEEE, 2012, pp. 533-536. doi: 10.1109/BCGIN.2012.145

[6] H. S. Cheraghchi, A. Zakerolhosseini, S. B. Shouraki, and E. Homayounvala, “A novel granular approach for detecting dynamic online communities in social network,” Soft Computing, 2018, pp. 1-22. doi: 10.1007/s00500-018-3585-z

[7] A. Moayedikia, “Multi-objective community detection algorithm with node importance analysis in attributed networks,” Applied Soft Computing, vol. 67, 2018, pp. 434-451. doi: 10.1016/j.asoc.2018.03.014

[8] U. K. Wiil, N. Memon, and P. Karampelas, “Detecting new trends in terrorist networks,” International Conference on Advances in Social Network Analysis and Mining, IEEE, 2010, pp. 435–440. doi: 10.1109/ASONAM.2010.73

[9] M. T. De Schotten et al. “A lateralized brain network for visuospatial attention,” Nature Neuroscience, vol. 14(10), 2011, pp. 1245–1246. doi: 10.1038/nn.2905

[10] W. Liu, M. Pellegrini, and X. Wang, “Detecting communities based on network topology,” Scientific Reports, vol. 4, 2014. doi: 10.1038/srep05739

[11] J. Yang and J. Leskovec, “Overlapping community detection at scale: a nonnegative matrix factorization approach,” The Sixth International Conference on Web Search and Data Mining, ACM, 2013, pp. 587–596. doi: 10.1145/2433396.2433471

[12] Z. Shi and A. B. Whinston, “Network Structure and Observational Learning: Evidence from a Location-Based Social Network,” Journal

of Management Information Systems, vol. 30(2), 2013, pp. 185–212. doi: 10.2753/MIS0742-1222300207

[13] C. Shi, Y. Cai, D. Fu, Y. Dong, and B. Wu, “A link clustering based overlapping community detection algorithm,” Data & Knowledge Engineering, vol. 87, 2013, pp. 394–404. doi: 10.1016/j.datak.2013.05.004

[14] T. Nepusz, A. Petróczi, L. Négyessy, and F. Bazsó, “Fuzzy communities and the concept of bridgeness in complex networks,” Physical Review E, vol. 77(1), 2008. doi: 10.1103/PhysRevE.77.016107

[15] G. Jia et al. “Community detection in social and biological networks using differential evolution,” In Learning and Intelligent Optimization, Springer, 2012, pp. 71–85. doi: 10.1007/978-3-642-34413-8_6

[16] Y. Y. Ahn, J. P. Bagrow, and S. Lehmann, “Link communities reveal multiscale complexity in networks,” Nature, vol. 466(7307), 2010, pp. 761–764. doi: 10.1038/nature09182

[17] Z. Pawlak, “Rough sets,” “International Journal of Computer & Information Sciences,” vol. 11(5), 1982, pp. 341–356. doi: 10.1007/BF01001956

[18] W. W. Zachary, “An information flow model for conflict and fission in small groups,” Journal of Anthropological Research, vol. 33(4), 1977, pp. 452–473. doi: 10.1086/jar.33.4.3629752

[19] B. Esteve-Altava, R. Diogo, C. Smith, J. C. Boughner, and D. Rasskin-Gutman, “Anatomical networks reveal the musculoskeletal modularity of the human head,” Scientific Reports, vol. 5, 2015. doi: 10.1038/srep08298

[20] D. Lusseau et al. “The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations,” Behavioral Ecology and Sociobiology, vol. 54(4), 2003, pp. 396–405. doi: 10.1007/s00265-003-0651-y

[21] J. Xie, S. Kelley, and B. K. Szymanski, “Overlapping community detection in networks: The state-of-the-art and comparative study,” ACM Computing Surveys, vol. 45(4), 2013, pp. 1–35. doi: 10.1145/2501654.2501657

[22] V. Krebs, “Books about US politics,” http://www.orgnet.com, 2004 (accessed February 1, 2009).

[23] M. Girvan and M. E. Newman, “Community structure in social and biological networks,” Proceedings of the National Academy of Sciences (PNAS), 2002, pp. 7821–7826. doi: 10.1073/pnas.122653799

[24] W. Liu, X., Jiang, M. Pellegrini, and X. Wang, “Discovering communities in complex networks by edge label propagation,” Scientific Reports, vol. 6, 2016. doi: 10.1038/srep22470

[25] A. V. Esquivel and M. Rosvall, “Comparing network covers using mutual information,”. arXiv Preprint arXiv:1202.0425, 2012.

[26] H. Shen, X. Cheng, K. Cai, and M.B. Hu, “Detect overlapping and hierarchical community structure in networks,” Physica A: Statistical Mechanics and Its Applications, vol. 388(8), 2009, pp. 1706–1712. doi: 10.1016/j.physa.2008.12.021

[27] S. Gregory, “Finding overlapping communities in networks by label propagation,” New Journal of Physics, vol. 12(10), 2010. doi:10.1088/1367-2630/12/10/103018

[28] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society,” Nature, vol. 435(7043), 2005, pp. 814–818. doi: 10.1038/nature03607

[29] C. Lee, F. Reid, A. McDaid, and N. Hurley, “Detecting highly overlapping community structure by greedy clique expansion,” Proceedings of the 4th Workshop on Social Network Mining and Analysis, 2010, pp. 33–42.

[30] K. R. Brown and I. Jurisica, “Online Predicted Human Interaction Database. Bioinformatics,” vol. 21(9), 2005, pp. 2076–2082. doi: 10.1093/bioinformatics/bti273

[31] Q. J. Jiao, Y. K. Zhang, L. N. Li, and H. B. Shen, “BinTree Seeking: A Novel Approach to Mine Both Bi-Sparse and Cohesive Modules in Protein Interaction Networks,” PLoS ONE, vol. 6(11), 2011. doi: 10.1371/journal.pone.0027646