AI & CYBER FORUM
AN INTERNATIONAL JOURNAL

OPEN ACCESS

ISSN 3117-6399
Original Research Article

Detecting Communities in Social Networks by Optimizing Similarity and Modularity Measures

Haritha Shree K. S.*, and Prerna Sidana

Glocal School of Technology and Computer Science, Glocal University, Saharanpur, Uttar Pradesh, India

Corresponding author: **Haritha Shree K. S.** | E-mail: **ksharitha.mbl@gmail.com**

Citation: Haritha Shree K. S., and Prerna Sidana (2024). Detecting Communities in Social Networks by Optimizing Similarity and Modularity Measures. *AI & Cyber Forum: An International Journal.* **DOI: https://doi.org/10.51470/AI.2024.3.1.01**

Received 05 January 2024 | Revised 03 February 2024 | Accepted 04 March 2024 | Available Online April 06, 2024

Abstract

This research presents a novel approach to revealing the community structure in big, complicated networks. The suggested approach for community detection makes advantage of the popular idea of optimizing network modularity. A cosine similarity measure that relies on shared links is utilized by our strategy to accomplish this. If you have a big network with a lot of nodes, this similarity metric will help you uncover their similarities quickly. Additionally, unlike other similarity metrics, it takes into account the network's sparse nodes, which are relatively simple. The approach then chooses the pairwise closeness among the complex network nodes after the similarity has been determined. It also identifies communities that are maximising modularity effectively by following a strategy inspired by the famous state-of-the-art Louvain approach. By conducting trials, we were able to illustrate that our strategy outperforms competing methodologies and yields somewhat better, more trustworthy outcomes. Methods' effectiveness in the network was assessed by looking at their modularity value, quality of community, and community size.

Keywords: *Social Networks; community network; Modularity and Similarity Measures.*

1. INTRODUCTION

The most common means of contact nowadays are social networks. People can voice their opinions, disseminate information, advertise products and services, publicize upcoming events, connect with powerful individuals, and engage in political discourse. Hence, data on the social media platform is growing substantially. The data explosion has presented fresh obstacles to the visualization and analysis of social network data. Networks with community structure as a key characteristic are common, and social networks are a good illustration of this. Finding groups within networks is an important problem. Consequently, numerous approaches to community detection have been put out in the literature thus far [1]. The behavior of complicated networks can be better understood with the aid of community structure. Similar to communities, it is a set of linked online pages around a certain subject. Communities aid in marketing, information transmission, comprehension, analysis, discovering hidden patterns, recommendation systems, and location-based interaction analysis. An example of a subgraph in a network is a community, which is a group of nodes (individuals) that facilitates communication and the growth of interpersonal relationships. Different names for communities include modules, groups, and clusters. According to [2] the community structure is characterized by a collection of nodes that are more densely packed inside a group and less densely packed between other groups.

This article introduces the "*Cosine Shared Link Method (CSLM)*" a novel technique that relies on cosine similarity. The foundation of CSLM is the cosine similarity measure and the agglomerative greedy technique. Analyzing real-world networks is possible using this technology, which is highly efficient.

2. Methodology

To identify communities in intricate networks, researchers have developed the Cosine Shared Link Method (CSLM) [3] a greedy agglomerative hierarchical approach. Starting from the ground up, CSLM uses a cosine similarity metric to repeatedly combine nodes into communities, and then uses modularity to assess the quality of these communities. There are three primary steps to the process:

Metric for Similarity: There is an initial assumption that each node is a separate community. The technique calculates the Cosine Shared Link Score (CSLS) for each node j and uses it to assess possible moves to each of its surrounding communities g, by calculating the cosine similarity and the density of their common connections. Once node j finds a neighbor with the highest CSLS, it shifts to that neighbor. Once the community structure has stabilized, the process is repeated for all nodes.

Optimization of Modularity: Following the establishment of initial communities, the algorithm calculates the network's modularity Q, a measure of the strength of the community structure, in order to optimize modularity. A function that determines the modularity is

$$Q = \frac{W}{2l} - \left(\frac{d_j \cdot d_k}{(2l)^2}\right)$$

where W is the weight of the link between nodes j and k_t d_j and d_k are the degrees of nodes j and k_r and l is the total number of connections in the network, respectively. The most coherent community structure is the goal of the algorithm, which seeks to maximize Q.

Aggregation of Community: Here, the algorithm takes the communities identified in the previous step and merges them. The network is reorganized such that each node is now considered as belonging to a larger community. The total link weights within each group are used to calculate the link weights among these aggregated communities. The combined network then undergoes the same procedure iteratively until the modularity value converges, indicating that the community structure has stabilized.
Using these processes iteratively, CSLM efficiently identifies community structures in large networks by leveraging modularity optimization and cosine similarity measurements to ensure accurate and coherent community detection.

3. Experiments and Results

The proposed "Cosine Shared Link Method (CSLM)" is tested on five real-world networks and compared with the PyLouvain and Louvain methods [4]. The evaluation criterion used is modularity [5]. Table 1.1 outlines the parameters used, and Table 1.2 provides descriptions of the datasets.

Table 1.1: An explanation of the parameters

Parameter	Meaning
n	Number of nodes
l	Number of links
Q	Modularity
C	Number of communities

Table 1.2: Datasets

Networks	Nodes	Links	Description
Football	115	613	Football network
Email	1133	10903	Email network
Facebook	4039	88234	Social circles from Facebook
Citations	27770	352807	Paper citations network
Brightkite	58228	214078	Location-based online social network

The Cosine Shared Link Method (CSLM) was tested against the PyLouvain and Louvain methods on five real-world networks (Football, Email, Facebook, Citations, and Brightkite). The comparison focused on modularity and community detection. CSLM consistently outperformed the other methods, achieving higher modularity values in most cases, as shown below in given tables and figures

Table 1.3: Modularity Score vs. Number of Nodes

Network	Number of Nodes	CSLM Modularity (Q)	PyLouvain Modularity (Q)	Louvain Modularity (Q)
Football	115	0.606	0.604	0.604
Email	1,133	0.567	0.564	0.566
Facebook	4,039	0.837	0.835	0.835
Citations	27,770	0.650	0.650	0.649
Brightkite	58,228	0.680	0.679	0.675

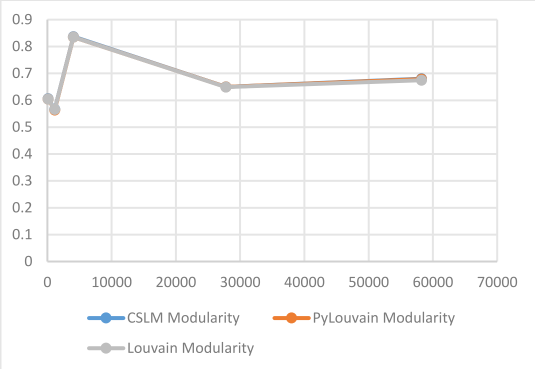

Fig. 1.3: Modularity Score vs. Number of Nodes

Fig. 1.3: Modularity score vs. number of nodes – This figure demonstrates that CSLM achieves higher modularity scores across networks with varying node sizes.

Table 1.4: Modularity score versus number of links

Network	Number of Links	CSLM Modularity (Q)	PyLouvain Modularity (Q)	Louvain Modularity (Q)
Football	613	0.606	0.604	0.604
Email	10,903	0.567	0.564	0.566
Facebook	88,234	0.837	0.835	0.835
Citations	352,807	0.650	0.650	0.649
Brightkite	214,078	0.680	0.679	0.675

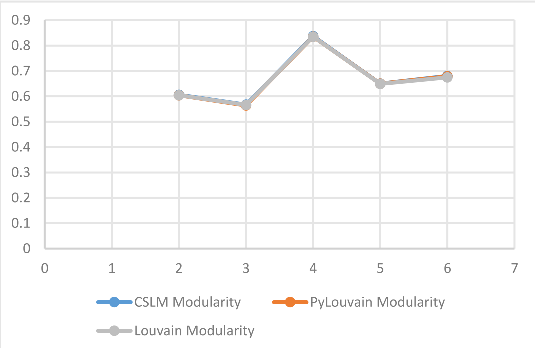

Fig. 1.5: Modularity score versus number of links

Fig. 1.4: Modularity score vs. number of links – It shows the modularity performance in relation to the number of links, again confirming CSLM's superiority.

Table 1.5: Number of communities versus number of nodes

Network	Number of Nodes	CSLM Communities (C)	PyLouvain Communities (C)	Louvain Communities (C)
Football	115	10	10	9
Email	1,133	12	12	10
Facebook	4,039	16	16	16
Citations	27,770	167	167	170
Brightkite	58,228	724	724	853

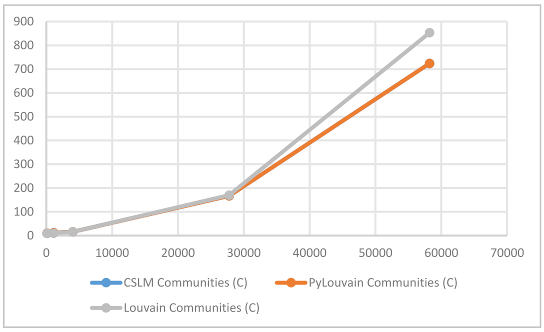

Fig. 1.5 Number of communities versus number of nodes

Fig. 1.5: Number of communities vs. number of nodes – CSLM detects a balanced number of communities, scaling appropriately with the network size.

Table 1.6: Number of communities versus number of links

Network	Number of Links	CSLM Communities (C)	PyLouvain Communities (C)	Louvain Communities (C)
Football	613	10	10	9
Email	10,903	12	12	10
Facebook	88,234	16	16	16
Citations	352,807	167	167	170
Brightkite	214,078	724	724	853

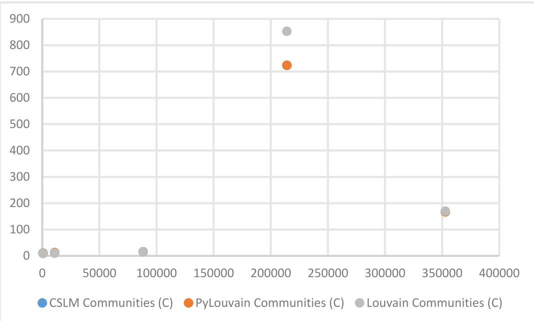

Fig. 1.6: Number of communities vs. number of links – It illustrates that CSLM performs better in detecting cohesive community structures as the number of links increases.

4. Conclusion

Identifying communities in big, complicated networks is a problem that has received a lot of attention recently. The community structure of complicated systems can be revealed using a number of recently developed efficient methods. Research into the examination of communities within intricate networks has recently attracted a lot of attention. The unveiling group in complicated networks has been the subject of a plethora of proposed methodologies. To reveal communities in complicated networks, we have presented a method, the CSLM approach, which is based on the iterative agglomerative greedy strategy. In complicated networks, the CSLM approach maximizes the modularity value. The value of modularity indicates the caliber of the network community. For this network community detection task, we have employed a novel similarity metric, CSLS. CSLS is a hybrid of the cosine similarity metric and the shared link metric. When two nodes are similar, the CSLS will find them. After that, the nodes that have a high CSLS are paired with the node that is most comparable to them. Moreover, CSLM's output is validated in actual networks. The CSLM achieves better results than the current methods for community detection. Finding partitions in big networks using the CSLM takes linear time. When compared to the current approaches, the modularity value produced with CSLM is significantly higher. This indicates that compared to other methods, the community structure produced by CSLM is of higher quality. The results of the trial validate the efficacy and precision of our method.

References

1. Morrison, M., & Gabbay, M. (2020). Community detectability and structural balance dynamics in signed networks. *Physical Review E*, *102*(1), 012304.
2. Riolo, M. A., & Newman, M. (2020). Consistency of community structure in complex networks. *Physical Review E*, *101*(5), 052306.
3. Chaudhary, L., & Singh, B. (2020). Community detection using maximizing modularity and similarity measures in social networks. In *Smart Systems and IoT: Innovations in Computing: Proceeding of SSIC 2019* (pp. 197-206). Springer Singapore.
4. Zhang, J., Fei, J., Song, X., & Feng, J. (2021). An improved Louvain algorithm for community detection. *Mathematical Problems in Engineering*, *2021*(1), 1485592.
5. Fang, Y., Huang, X., Qin, L., Zhang, Y., Zhang, W., Cheng, R., & Lin, X. (2020). A survey of community search over big graphs. *The VLDB Journal*, *29*, 353-392.