# A NOVEL METHOD FOR DETECTION AND ANALYSIS IN A NETWORK OF OVERLAP ABLE COMMUNITIES

**Bashar Mohammed Tuama - Computer Science - Altinbas university**

**Turkey Istanbul- Ministry of Health / Salah Al-Din Health Department / Balad Sector**

**Laith Ali Abdulsahib - department of Computer science- faculty of education for women- university of kufa- Iraq**

laith.albaldawi@uokufa.edu.iq

## Abstract:

In the domain of Network Data Processing and Data Learning, the identification of communities is crucial for comprehending the functional characteristics of networks. Overlapping community detection, which involves clusters with shared nodes, has become increasingly important in the context of real-world networks. However, there is still a requirement for additional research and the creation of innovative algorithms that consider various factors.

This research proposes an updated method for perceiving the inferences within network communities. It introduces a multi-stage approach that starts by identifying seed nodes and concludes by discovering overlapping communities. The novelty lies in the use of a graph/network metric to identify significant seed nodes. The research focuses on two categories: identifying highly significant nodes based on similarity measures and recognizing cluster centers that maximize community density.

The experimental outcomes validate the efficiency and scalability of the proposed methodology in identifying overlapping communities in large-scale real-world networks. Through a comparative analysis against state-of-the-art methods, the performance of the proposed approach is further confirmed.

This study makes a significant contribution to the field of community detection by presenting an innovative approach that takes into account overlapping communities and integrates graph/network metrics. The results offer valuable insights into the characteristics and functional properties of networks, thereby facilitating advancements in network data processing and data learning methodologies.

Keywords: (Overlapping Community Detection, Seed Algorithm, Facebook, Data Learning, Network Metric).

## I. Introduction

### A. Background

In the realm of network data processing and data learning, the detection of communities plays a crucial role in understanding the functional properties of various types of networks. A community, also known as a cluster, refers to a set of vertices that are densely interconnected within themselves and exhibit significant connections to other communities within the network. Recognizing the community structure provides valuable insights into the network's behavior and operational characteristics. While early research primarily focused on disjoint or non-overlapping community detection, it has become evident that real-world networks often exhibit overlapping community structures, where nodes can belong to multiple communities simultaneously. This shift in focus has led to the emergence of overlapping community detection as an important research area.

### B. Problem Statement

The existing literature offers various algorithms for community detection; however, there is still ample room for further investigation and the development of novel algorithms that account for different factors. In this context, this research aims to address the need for an up-to-date method to perceive the inferences that occur within network communities. The primary objective is to develop a multi-stage approach that starts by identifying seed nodes and culminates in the discovery of overlapping communities within the network. Unlike traditional approaches, this method incorporates a graph/network metric to identify the seed nodes, enabling a more effective and comprehensive community detection process.

### C. Research Goal

The primary objective of this research is to introduce a new method for detecting overlapping communities in networks. The research aims to develop a multi-stage approach that starts by identifying seed nodes and then proceeds to uncover overlapping communities within the network. This approach utilizes graph/network metrics to identify important nodes as seed nodes and takes into account both similarity measures and community density metrics. The effectiveness and scalability of the proposed approach will be evaluated using real-world networks, and its performance will be compared to existing state-of-the-art algorithms for overlapping community detection.

This research aims to contribute significantly to network data processing by providing valuable insights into the characteristics and functional aspects of overlapping communities in diverse

network types. By achieving these objectives, it seeks to enhance our understanding of network dynamics and improve data processing techniques in various domains.

## II. Literature Review

Theoretical background plays a crucial role in understanding the interconnections and identification methods within networks. Local neighborhood structures and their algorithmic approaches have been extensively studied in the literature. Previous studies, such as [1], have focused on representing and analyzing local neighborhood structures in networks.

In [2], a comprehensive review of various local community detection algorithms is provided. The analysis compares multiple algorithms and identifies the variance among them. The evaluation is based on extensive testing using benchmark networks with varying distribution of communities.

The problem of community detection can be visualized as a local problem within a network, and graphs are the preferred data format for network analysis. [3] discusses different techniques for graph clustering and their performance. These techniques help address the challenges of identifying neighborhoods within a network.

Research has been conducted on clustering and community detection in coordinated networks [4]. This research focuses on networks with coordinated and cutting-edge structures, exploring various active systems for community identification. Metrics and evaluation methods are provided, along with suggestions for future research directions.

The identification of overlapping communities has gained significant attention due to the prevalence of overlapping characteristics in large-scale networks. [5] examines existing algorithms for identifying overlapping communities and considers benchmark datasets and quality criteria. The evaluation highlights the differences between performance metrics at the node and community levels.

Different classes of algorithms have been developed for identifying overlapping communities. Class 1 includes algorithms designed to handle large-scale datasets, such as [6][7][8]. These algorithms use stochastic slope, Markov chain Monte Carlo methods, and other techniques to efficiently locate community centers.

Class 2 algorithms focus on the topological characteristics of graphs, such as cliques, related components, and edge loads. For example, [9] proposes a strategy for neighborhood identification based on hub combinations and uses the PageRank method to rank nodes in the global network.

Class 3 algorithms, known as seed-driven methods, are particularly suitable for identifying overlapping communities. These methods involve expanding local communities and comparing them to global objective-based partitioning of networks. Examples include [10]-[13]. The proposed methodology in this study falls within this class.

The selection of appropriate network/graph metrics is crucial for community detection. [14] provides a comprehensive overview of network measurements, categorizing and examining their significance. [15] categories and groups various graph measurements and describes their relationships, aiding in the selection of relevant measurements.

Different network/graph analysis tools and libraries are available, each with its own capabilities and specialties. Researchers can consult [15] for options to determine which tool to use for their specific calculations. Additionally, libraries like JUNG provide a unified and extensible framework for modeling, analyzing, and visualizing graph data using Java [16].

In summary, the theoretical background covers a range of topics related to interconnections and identification methods in networks. The literature review provides insights into local neighborhood structures, different classes of algorithms for community detection, network/graph measurements, and analysis tools. This knowledge serves as the foundation for the proposed algorithm and its implementation.

III. Methodology

Designing and Developing the Problem

The goal of community detection is to identify the location of community C1, whereas traditional network clustering faces challenges and aims to divide the network into k distinct groups C1, C2, and Ck, satisfying the condition C1 U C2... U Ck = V. Unlike traditional clustering, where Ck encompasses all vertices, overlapping communities may not include all vertices of the graph, i.e., C1 U C2....U C2...Ck.

Flow of Steps in the Proposed Method: The proposed method consists of five phases as shown in Figure 1. In Phase 1 of the proposed method, the focus is on identifying the steps for bi-connected clusters. The following methods are employed:

i. The largest connected subgraph of the input graph G = (VC, EC), denoted as the Biconnected Core GC = (VC, EC) (V, E\Es), is determined. This subgraph represents the core structure of bi-connected clusters.

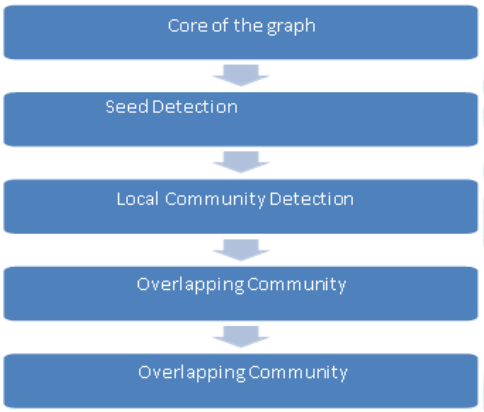


Figure 1: Shows the steps of the proposed Method

ii. Before calculating the doubly connected kernel, subgraphs connected to the bi-connected kernel of the input graph are removed. This step helps streamline the identification process.

By excluding nodes that do not contribute to the community structure, the algorithm eliminates outliers and unnecessary components from the graph. These nodes, which are deemed irrelevant, will be reintroduced in the final phase to restore the original graph structure. The doubly connected kernel of a graph represents the subgraph with the highest level of connectivity within the induced subgraph. The existence of a substantial number of edges within this doubly connected core is essential for the presence of the desired communities.

Phase 2 of the algorithm focuses on the selection of seed nodes for the community detection process. It is a crucial step that significantly impacts the overall performance and accuracy of the algorithm. starts by setting D, which represents the graph's vertices arranged in descending order. The seed number k is determined using sample identification or correct ability analysis.

In this phase, the seed set S is generated as the output, initially starting as an empty set. The algorithm proceeds by iterating through each vertex x in D and evaluating its Similarity Central importance (BC) in relation to all its unlabeled nearest points.

If the BC(x) value exceeds the threshold, indicating the significance of the vertex, it is added to the seed set S. The vertex x is also labeled as a visited node. This process continues until all vertices in D have been processed. There are two ways to determine the value of k, representing the selected starting seed nodes. One approach is to use the ground truth information from the network, while the other approach involves utilizing the seclusion report produced by the instrument and applying a clustering method to determine the number of modules or communities in the network.

In phase 3 of the algorithm, the focus shifts to the identification of local intersecting groups, utilizing the seed nodes obtained from the previous step. The process begins by creating local communities (LC) for each seed node (s) in the seed set (S). This is achieved by combining the seed node (s) with its corresponding neighborhood set (NS). The algorithm then iterates through each seed node in the seed set, calculating the local set (NS) by employing an appropriate algorithm or method to determine the neighboring vertices or nodes of the seed node. Once the local communities have been established, the process is finalized.

This phase is instrumental in uncovering the communities present within the network, as it effectively captures the interplay between the seed nodes and their respective neighborhoods. By creating these local intersecting groups, the algorithm provides valuable insights into the community structure of the network. This deeper understanding of the relationships and connections between different nodes enhances our knowledge of the network's functional properties and facilitates further analysis and interpretation of the data.

Phase 4 of the algorithm focuses on the identification of global communities based on the local communities discovered in the previous phase. The process starts by leveraging the local communities (LC) obtained from the previous phase and the network diagram (G) to discover the global communities. The output of this phase includes two types of communities: communities with sparse overlap (SOC) and communities with a lot of overlap (DOC). To begin, the vertices that do not belong to the local community set (LC) are labeled as LC'. Next, for each local community lc in LC, the leaf node set LN(lc) is calculated, which serves as a crucial step in the algorithm. This step involves determining the distance between each pair of nodes in the leaf node set from LC' to every node in LC, denoted as (i, j). If the distance between nodes i and j falls below the specified limit for the similarity measure, they are added to the sparse overlap set (SOC). This process is repeated for all pairs of nodes. The algorithm pays special attention to the densely intersecting groups identified by the DOC sets, as they offer valuable insights into the various communities present in the network. By considering the

overlap between communities, the algorithm provides a comprehensive understanding of the community structure, enabling researchers to gain meaningful knowledge about the interconnections and relationships within the network.

IV. Datasets

We conducted experiments using different real-world networks to evaluate the accuracy of each algorithm. A benchmark dataset was employed to assess the performance of the new algorithm. [18] [19].

| Dataset Name | Number of Nodes | Number of Edges | Category of Dataset |
|---|---|---|---|
| Local club | 34 | 78 | Social |
| Dolphin Social Network | 62 | 159 | Social |
| . College Football | 115 | 616 | Social |
| Facebook | 4,039 | 88,234 | Social |

v. Experimental Result

In the first stage, the analysis of the information network was conducted (Figure 2A). In the second stage, the network was transformed, as shown in Figure 2B, which represents the initial filtering step using the Dolphin Social Network. The selection of seed nodes was based on the betweenness centrality score, with Figure 1C illustrating the distribution of these scores for the nodes in the Dolphin network. The top four values in the distribution were chosen as the seed nodes, as they exhibited higher betweenness and centrality scores.

The Dolphin social network, representing the network after stage 3, is depicted in Figure 2C, highlighting the local communities formed around the seed nodes. These local communities consist of nodes in close proximity to the seed nodes, and their growth is facilitated by the overlapping nature of the network.

Figures 2D, 2E, and 2H display pairs of communities with significant overlap, while Figure 12G shows a pair of communities with minimal overlap. Blue and green nodes represent distinct communities, with pink nodes serving as bridges between these communities. The black nodes not only act as endpoints between communities but also function as seed nodes for the network.

From these observations, we deduce the following regarding the metric used:

i) Seed nodes do not always hold the primary spotlight within a community.

ii) Their significance lies in their ability to facilitate substantial overlap between communities and serve as bridges that connect different networks.

iii) Our metric effectively determines crucial overlaps and classifies networks as dense or sparse based on predefined thresholds.

iv) Hence, the choice of using Betweenness Centrality as the metric was appropriate.

Given that community overlap is the primary feature of interest in this study, communities with minimal overlap are considered less important for analysis. Consequently, during the integration in stage 4, the connectivity exhibits a grouping of closely packed and sparsely overlapping communities.

Following the completion of stage 5, the Dolphin Social Network serves as the representation of the network. The filtered duplex components are incorporated into the duplex core to restore the original network structure. The number of graphs and edges accurately represents the total count of nodes and edges within the network. It is important to note that this post-processing step does not impact the classification of overlapping communities or the identification of distinct nodes.

At this stage, the algorithms have converged, successfully uncovering the overlapping communities within the Dolphin Social Network. Furthermore, experiments were conducted using Facebook's dataset, and a statistical summary of the calculations is presented.

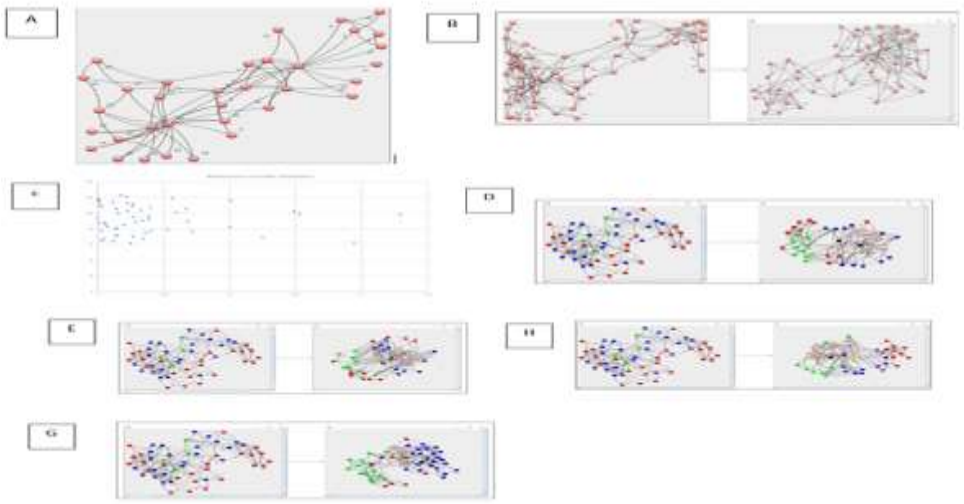

Figure 2: STEPS OF THE PROCEEDING OF GROUP NETWORKING

VI. Discussion

SUMMARY OF THE OBTAINED INTERLEAVED NETWORK

The analysis of the Dolphin social network revealed that it consists of numerous dense k-1 networks, with substantial overlap that can be considered for specific purposes. The results indicate that the selected seed nodes play crucial roles in larger overlapping communities. These seed nodes exhibit high significance and practical value. Nearly every vertex in the network is connected to a local community, resulting in extensive coverage and only a few unassigned nodes. Similar observations were made for the Facebook community, where overlapping networks help identify key communities that facilitate information dissemination and capture various user activities.

EVALUATION OF OUR PROPOSED METHOD AGAINST OTHER STATE-OF-THE-ART METHODS

Comparing the distance centrality metric with the betweenness centrality metric revealed notable differences. Distance centrality, which measures the proximity of a node to other vertices in the graph, did not identify significant features within nodes as effectively as our proposed metric. The specificity of the detected communities was better captured by our metric. Figure 4 demonstrates how the observed level of overlap strongly influences the specificity of the networks. Figures 3 and 4 illustrate the network structure after stages 2 and 3, while Figure 4 represents the network after stage 4 of the algorithm.

In comparison to other existing methods, our approach emerged as a superior performer. When evaluating our technique against other methods, we found that the results were generally equivalent. Therefore, we argue that our approach surpasses traditional methodologies in the same field. The chosen metrics proved to be valid and valuable at all stages of the computation, effectively identifying networks with high coverage, particularly overlapping associations. The coverage rate further demonstrates that the algorithm can effectively cover the entire network and accurately group communities. In addition to providing a comprehensive quantitative analysis, we also present two categories of overlapping communities: sparse overlap and dense overlap. This demonstrates the compatibility of our algorithm with previous studies and established methodologies.

Overall, the obtained interleaved network analysis and the evaluation of our proposed method validate its effectiveness in detecting and characterizing overlapping communities within complex networks.

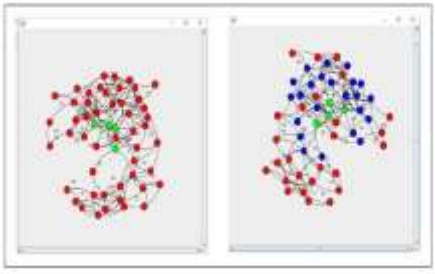

Figure 4: Employing the Length Connectivity Criterion Steps 1 and 2

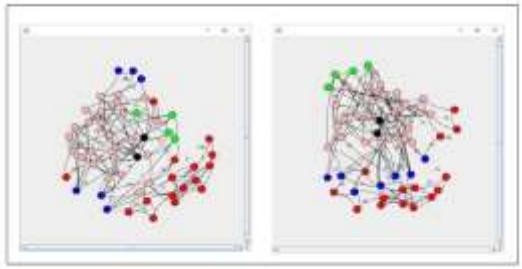

Figure 3: Employing the Length Connectivity Criterion Steps 1 and 2

VII. Conclusion

In this work, we have focused on developing an innovative seed-driven approach to address the challenge of identifying influential network hubs. We have recognized the significance of this problem and its relevance in various real-world scenarios. Through a comprehensive analysis of different graphs and network metrics, we have established the importance of our chosen metric. Our proposed algorithm, which consists of multiple steps, effectively captures the entire process of identifying overlapping communities in a computationally efficient manner. Each step contributes to the overall framework and holds its own significance.

Our contributions are twofold. Firstly, we employ a seed-based strategy to identify the key hubs in the network, allowing for a targeted exploration of influential nodes. Secondly, our approach accommodates both dense and sparsely connected networks, without compromising on local community segregation. Importantly, our method can be applied to large-scale real-world networks, where we successfully maintain crucial community attributes such as isolation, cohesion, and coverage.

We have compared our algorithm to existing state-of-the-art methods, particularly focusing on the distance centrality measure. Our results demonstrate that our algorithm outperforms other techniques, including distance centrality, in determining the quality of overlapping networks. This suggests that the communities identified by our algorithm possess significant practical relevance within the network, owing to their overlapping properties.

In conclusion, our work presents an original and effective approach to address the challenge of identifying influential network hubs. Our algorithm considers the covering properties of communities and successfully captures important aspects of the network's local community structure. This research has implications for various domains where understanding community interactions and influential hubs is crucial for decision-making and problem-solving.

VII. References

Girvan, M., and M. E. J. Newman. Community Structure in Social and Biological Networks. Proceedings of the National Academy of Sciences of the United States of America 99.12 (2002): 7821–7826. PMC. Web. 5 Dec. 2016.

Lancichinetti, Andrea, and Santo Fortunato. Community detection algorithms: a comparative analysis. Physical review E 80.5 (2009): 056117.

Schaeffer, Satu Elisa. Graph clustering. Computer science review 1.1 (2007): 27-64.

Malliaros, Fragkiskos D., and Michalis Vazirgiannis. Clustering and community detection in directed networks: A survey. Physics Reports 533.4 (2013): 95-142.

Xie, Jierui, Stephen Kelley, and Boleslaw K. Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. ACM Computing Surveys (csur) 45.4 (2013): 43.

Gopalan, Prem K., and David M. Blei. Efficient discovery of overlapping communities in massive networks. Proceedings of the National Academy of Sciences 110.36 (2013): 14534-14539.

El-Helw, Ismail, Rutger Hofman, and Henri E. Bal. Towards Fast Overlapping Community Detection. Cluster, Cloud and Grid Computing (CCGrid), 2016 16th IEEE/ACM International Symposium on. IEEE, 2016.

El-Helw, Ismail, et al. Scalable Overlapping Community Detection.

Wen, Xuyun, et al. A Maximal Clique Based Multiobjective Evolutionary Algorithm for Overlapping Community Detection. IEEE Transactions on Evolutionary Computation (2016).

Hu, Yanmei, et al. Voting based seeding algorithm for overlapping community detection. Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2015 International Conference on. IEEE, 2015.

Moradi, Farnaz, Tomas Olovsson, and Philippas Tsigas. A local seed selection algorithm for overlapping community detection. Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on. IEEE, 2014.

Xu, Bingying, et al. Local community detection using seeds expansion. Conference Anthology, IEEE. IEEE, 2013.

Lee, Conrad, et al. Seeding for pervasively overlapping communities. Physical Review E 83.6 (2011): 066107.

Costa, L. da F., et al. Characterization of complex networks: A survey of measurements. Advances in physics 56.1 (2007): 167 242.

Hernández, Javier Martın, and Piet Van Mieghem. Classification of graph metrics. Delft University of Technology, Tech. Rep (2011).

http://www.kdnuggets.com/2015/06/top-30-social-network-analysis-visualization-tools.html

http://jung.sourceforge.net/

D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, The Bottlenose dolphin community of Doubtful Sound features a large proportion of long- lasting associations, Behavioral Ecology and Sociobiology 54, 396-405 (2003)

J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.