Amrita School of Computing
Department of Computer Science and Engineering

B.Tech 2020-24 CSE- Project Phase 1

Minor Project: 19CSE495

Proposal

Group C4

I. **Project Title:**
Overlapping Community Detection on Attributed Networks.

II. **Team members:**

Roll No.	Name
AM.EN.U4CSE20275	Vennam Venkata Siva Naga Sai Mohan
AM.EN.U4CSE20217	Boora Jagadananda Tharaka
AM.EN.U4CSE20248	Mudiyala Vishal Reddy
AM.EN.U4CSE20215	Balla Hemanth

III. **Abstract**

To address the problem of identifying overlapping communities in networks while considering node attributes as well as the topological structure of the network. In many real-world scenarios, such as social networks or biological systems, nodes can belong to multiple communities simultaneously. The motivation here is, general community detection methods ignore the presence of overlapping communities and fail to consider the rich information provided by node attributes. By combining the information from node attributes and the network topology, this project aims to develop an algorithm that can effectively detect overlapping communities. This problem is relevant as it has applications in various domains, including social network analysis, and recommendation systems. Challenges include developing efficient algorithms that consider attribute and structural interactions and handling large-scale attributed networks with diverse attribute types.

IV. Background Study

Title &year	Problem	Contributions	Limitations	Open problems/Future work
Ge, Jinhuan, et al. "LPX: Overlapping community detection based on X-means and label propagation algorithm in attributed networks." *Computational Intelligence* 37.1 (2021): 484-510.	Detecting the community structure of a graph only through structural information or attribute information is not an effective approach. A person may change his or her role in different groups at any time depending on the environment in a social network.	This algorithm can identify overlapping communities in an attributed network based on the rule of theme weight. Our algorithm performs redundant processing to avoid community redundancy. This algorithm does not require the user to specify other parameters.	The user needs to specify the number of clusters in advance, which is a difficult task. The clustering result is the local optimum. LPA treats all nodes in the network equally and updates the node label with a random sequence. The algorithm considers all neighboring nodes of a node to have the same influence.	Although the LPX algorithm can effectively detect the community structure in the attributed network, further improvement is needed. The next step is to improve the LPX algorithm to address the uneven number of nodes and number of attributes in the attributed network.
Malhotra, Deepanshu, and Anuradha Chug. "A modified label propagation algorithm for community detection in attributed networks." *International Journal of Information Management Data Insights* 1.2 (2021): 100030.	One of the fundamental problems consists of finding the cohesive clusters or dense groups of nodes, known as communities, in such systems. The detection of such communities is crucial for various other tasks like recommendation, marketing, distribution, etc.	The authors propose a modified LPA that uses node attributes and link strength measures to overcome the random selection problem in community detection. This paper depicts about four variants of their method that can handle different types of node attributes and have near linear time complexity.	1)Sensitivity to Initi alization: To give e ach node a distinct c ommunity label, the proposed technique uses rand om initialization. 2) Paper provides a comparative analysis of the proposed variants with classical and node-attributed algorithms.	Extending the current paradigm to recognise overlapping communities where e ach node can connect to multiple communities, like in r eal- world systems. In rea l- world situations hom ogeneity is not guaranteed therefore, the need to solve problems in het erogeneous systems is increasing.

He, Chaobo, et al. "Semi-supervised overlapping community detection in attributed graph with graph convolutional autoencoder." *Information Sciences* 608 (2022): 1464-1479.	The existing methods for community detection in the attributed graph still cannot well solve three key problems simultaneously: 1. link information and attribute information fusion 2. prior information integration 3. overlapping community detection.	The paper introduces a semi-supervised graph convolutional autoencoder (SSGCAE) model for the task of overlapping community detection in attributed graphs. SSGCAE is designed to effectively integrate prior information and fuse link and attribute information, leading to improved community detection accuracy.	1. The proposed SSGCAE model on a relatively small number of real attributed graphs and synthetic datasets. Evaluations on a wider range of datasets would strengthen the generalizability 2. The modularity maximization module used in SSGCAE can be computationally expensive, the complexity might become a limitation.	1. Handling Noise: Future work could focus on developing techniques to handle noisy or incomplete attribute data in attributed graphs while handling real-world datasets. 2. Scalability and Efficiency Improvements: Might face scalability challenges on large graphs because of computation. Better graph convolutional architectures computing strategies to make the model more scalable.
Lin, Hanyang, et al. "Overlapping Community Detection Based on Attribute Augmented Graph." *Entropy* 23.6 (2021): 680.	The existing methods for community detection in the attributed graph still cannot • efficient link-attribute fusion • robust prior information integration • scalability • interpretability of detected communities in attributed graphs.	1. OCEA: Introduces innovative approach for community detection with attribute information using augmented attribute graph and fuzzy k-medoids. 2. AOCEA: Extends OCEA by adding a step to estimate the number of communities for improved detection. 3. Evaluated on real-world and synthetic networks: outperformed baselines, robust and accurate in detecting overlapping communities.	1. Attribute noise limits the accuracy of OCEA and AOCEA on networks. 2. AOCEA runtime escalates with large networks, impacting scalability. 3. AOCEA struggles with sparse community structures and identifying overlapping communities.	1. Improve estimation in AOCEA for accurate community count. 2. Use biased random walk for high attribute weight vertices in sparse networks. 3. Implement approx. random walk to enhance efficiency in detecting overlapping communities on massive datasets. 4. Future work on handling attribute noise, scalability, and improving accuracy in sparse networks.

Teng, Xiangyi, Jing Liu, and Mingming Li. "Overlapping community detection in directed and undirected attributed networks using a multiobjective evolutionary algorithm." *IEEE transactions on cybernetics* 51.1 (2019): 138-150.	Overlapping community detection using attribute information and structural connections synchronously. Especially, for the analysis of both directed and undirected networks.	1) Designed a new objective called EQOV to deal with directed and undirected networks. 2) Proposed a novel two-part encoding and decoding strategy to represent overlapping communities of single individuals without assigning the number of communities in advance. 3) Conducted experiments on both synthetic networks as well as directed and undirected real-world attributed networks.	1) The paper does not discuss the scalability of MOEA-SAOV in terms of large-scale networks. Community detection in large networks can be computationally intensive, and it is important to evaluate. 2) The paper does not provide information about the time and space complexity of MOEA-SAOV.	1) Improve the algorithm which can be suitable for large networks. Increase the scalability of MOEA. 2) Implementing the algorithm in the parallel pattern and using the surrogate model as well as the techniques from the estimation of the distribution algorithms to reduce the computational time in detecting community structures of networks with thousands of nodes.

V. Challenges

1) Scalability: Overlapping community detection in large-scale networks poses a significant challenge due to the increased complexity and computation required to identify multiple community memberships for nodes.
2) Resolution Limit: The resolution limit problem occurs when the algorithm fails to detect smaller, tightly knit communities within larger ones, leading to inaccurate results and loss of granularity.
3) Noise and Ambiguity: Real-world networks often contain noise and ambiguous connections, making it difficult to precisely determine the boundaries of overlapping communities and potentially leading to misinterpretations of community structures.

VI. Deliverables of Phase I

Outcomes/Deliverables

- Data Preprocessing: Develop a data preprocessing pipeline to handle noise and optimize the network representation for accurate community detection.
- Algorithm Selection: Evaluate and select promising algorithms or adapt existing ones to address the scalability and resolution limit challenges specific to the project.
- Attributes and Topology: First, implement algorithms using attributes and topology separately, and then fuse both attribute and topology.

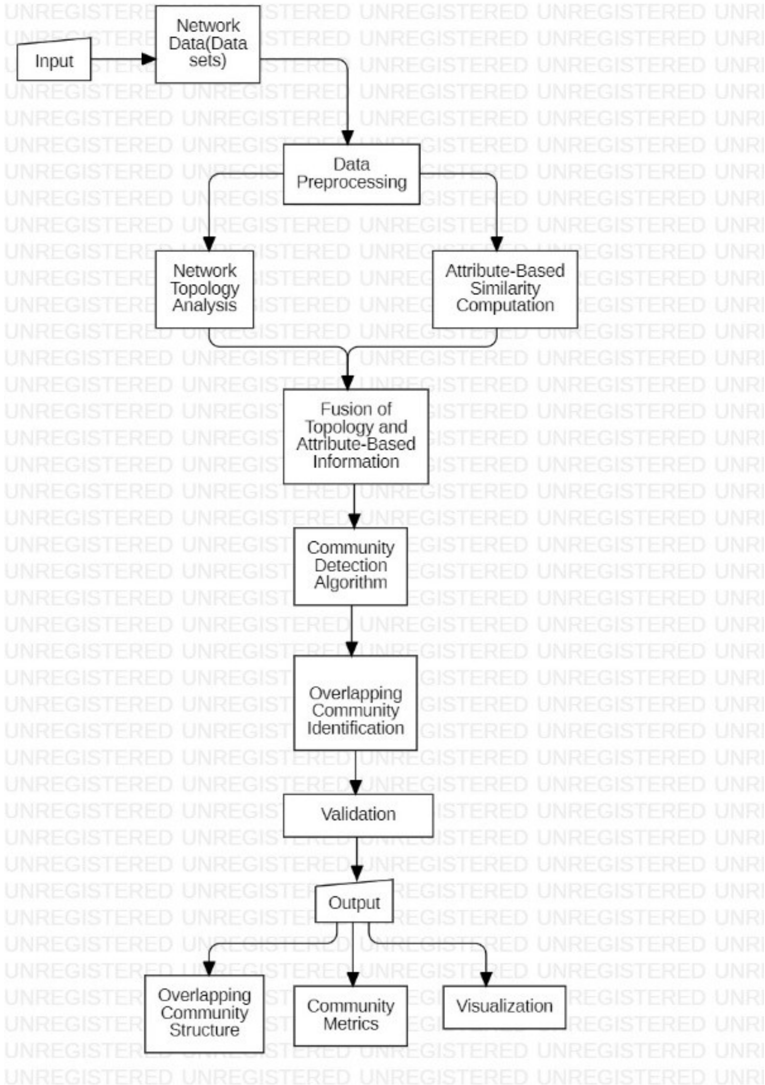
VII. Assumptions/Declarations:

The datasets should include both network structure and attribute information about the nodes so that they can be combined together for community detection.

VIII. Tools to be used

Software/Hardware Tools	Specifications
Python	We are using libraries such as networkx, igraph, gephi and cytoscype.

IX. High Level Design



Vennam Venkata Siva Naga Sai Mohan:

Boora Jagadananda Tharaka: B. Jagadanand

Balla Hemanth: Hemanth

Mudiyala Vishal Reddy: M- Vishal Reddy

Ms. Deepthi LR (Project Guide): Deepthi