巍巍天山雪， 泱泱红湖浪
高山人仰止， 源远水流长

新疆大学
XINJIANG UNIVERSITY

Research on Dynamic Community Detection Method Based on Multi-dimensional Feature Information of Community Network

College：School of Computer Science and Technology
Name：Hu Kui
Research direction：Dynamic Community Detection
Guidance teacher：Zhang Zhenyu、Li Xiaoming

新疆大学
XINJIANG UNIVERSITY

Self introduction

Hello everyone, I am Hu Kui, a master's student at the School of Computer Science and Technology, Xinjiang University. Currently, I am in my third year of graduate studies. My main research direction is to conduct related research on dynamic communities, such as dynamic community detection and community evolution prediction. My paper was completed under the guidance and assistance of Professor Zhang Zhenyu from Xinjiang University and Professor Li Xiaoming from Zhejiang Yuexiu University.

巍巍天山雪，泱泱红湖浪
高山人仰止，源远水流长

新疆大学
XINJIANG UNIVERSITY

The significance of research

With the passage of time, the relationships between clubs in the network are not static, but exhibit dynamic and changing characteristics. Studying dynamic community detection can provide insights into the inherent laws of network evolution. This not only provides strong support for predicting the future development trend of the network, but also lays a solid foundation for identifying key nodes in the network. In addition, dynamic community detection has shown broad application prospects in multiple fields, such as social network analysis, bioinformatics, and financial market modeling, providing powerful tools and ideas for solving practical problems and promoting the development of related fields. The club structure of Karate Club Karate Network is shown in Figure 1.

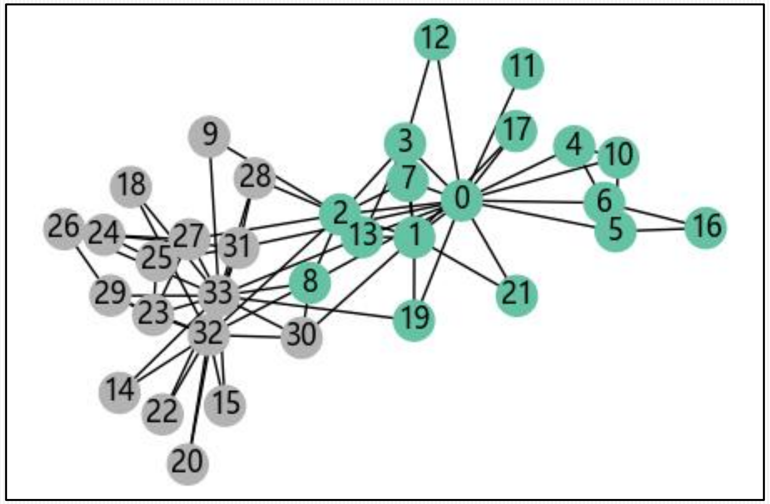

Figure 1

Problem Description

In the existing community detection methods, most methods for processing a single time snapshot can be roughly divided into clustering based, modularity based, and complete subgraph based methods. These methods mainly focus on the topological structure and degree centrality of the graph, and do not fully utilize the feature information of nodes. Taking e-commerce shopping platforms as an example, all users on the platform have become a community network that changes over time. Each user represents a node in the network, and users who are interested in the same type of product are divided into one category. Each user is interested in different products at different times, thus being grouped together with different users. In this process, the basic information of each user and the type of products they like (such as favorites, purchases, and recommendation behaviors) constitute the node feature information of the user at a certain moment in the dynamic community. The association between the user and their belonging to the same category constitutes the topology structure of the dynamic community at that moment.

Problem Description

Dynamic community detection needs to consider the constantly changing community structure, including the appearance and disappearance of nodes, edges, etc., that is, to consider the evolutionary information of dynamic communities. The current method mainly introduces the traditional static community detection method into temporal analysis to achieve the detection of dynamic communities. Although incremental clustering methods can utilize historical information, they ignore the feature information of nodes and are prone to getting stuck in local optima, resulting in a decrease in accuracy when the community structure undergoes significant changes.

Overall, the main problems with current dynamic community detection methods are the ineffective utilization of feature information from graph nodes or edges, as well as the inability to effectively combine historical information for data processing, which affects the accuracy of community detection structures.

Model Introduction

This article proposes a dynamic community detection method based on multi-dimensional feature information of communities(Dcdmf). The data processing method of this model involves integrating the node features and edge information of a graph into a Data object, which is then input into the algorithm for further processing.

The Dcdmf model will convert the node feature information in the pending community snapshot into a feature matrix before running, and the topology structure of the community will be transformed into an adjacency table. In the feature processing stage, the model will obtain representation information of all nodes within a range of 2 path lengths starting from each node. This processing method not only considers the characteristics of the node itself, but also integrates the information of its neighboring nodes, making the description of node features more comprehensive and rich. Next, the model will refer to historical community information, finely adjust the current community information, and integrate the information from this snapshot into the hidden state and cell state, making it easier to use in the next snapshot. At the same time, in order to facilitate subsequent community partitioning, the model will also scale the length of the node vector to match the number of communities and output a representation matrix. Finally, the nodes are divided into corresponding sub communities based on the representation matrix output by the model. The model diagram is shown in Figure 2.

巍巍天山雪，泱泱红湖浪
高山人仰止，源远水流长

新疆大学
XINJIANG UNIVERSITY

Model Introduction

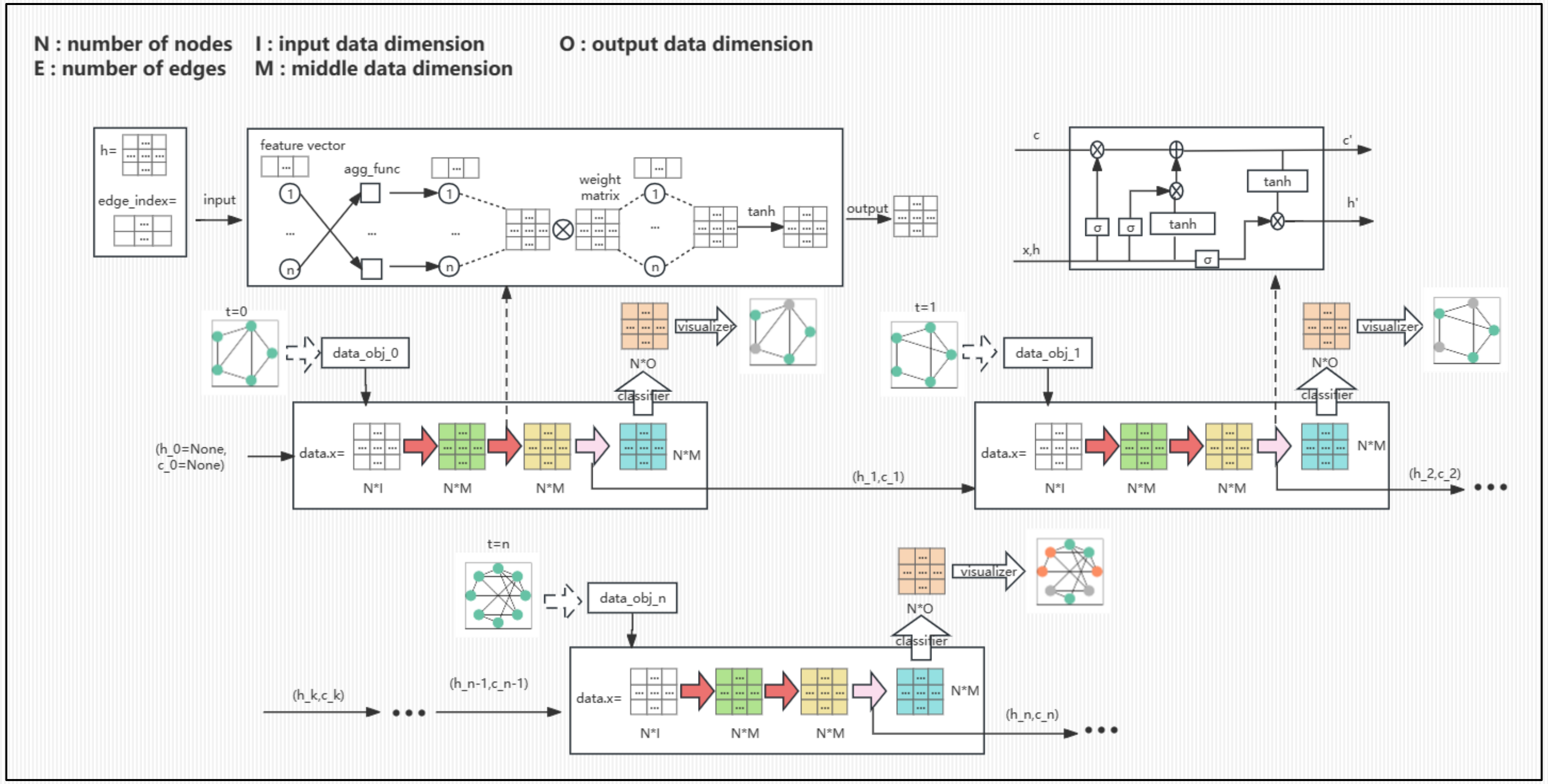

Figure2:Dcdmf model diagram

Datasets

1. Synthetic Dataset: It contains 3 snapshots, each containing a matrix A1 of 5 types of nodes. Use random methods to rewire the edges. Connect one set of edges to another based on the set probability. In this dataset, all link matrices (A) have 5000 nodes and 20000 edges. For the content matrix (C), generate five sets of five words and the probability of one word in multiple sets. Due to its structural nature, all content matrices have 25 words. The probability of random rewiring is p=0.75, and the probability of words appearing in a group is q=0.25.

2. DGraphFin Dataset (https://dgraph.xinye.com/dataset): It is a directed, unweighted dynamic graph composed of millions of nodes and edges, containing four types of users. It represents the social network between users of the Finvolution Group, where nodes represent Finvolution users, and the edge from one user to another means that users consider another user as an emergency contact. In addition, DGraph Fin based on real graph data can effectively promote the observation and understanding of evolutionary social networks. In addition, dynamic graphs can be adaptively applied to accurately identify fraudsters by detecting anomalies, which helps to improve the efficiency and effectiveness of financial risk control. Below is an explanatory overview of the dataset. It provides 821 snapshots (by talent), 3700500 nodes, and 4300999 edges.

巍巍天山雪，泱泱红湖浪
高山人仰止，源远水流长

新疆大学
XINJIANG UNIVERSITY

Datasets

Table 1: Dataset statistics

名称	节点数	边数	特征	快照数	类别	方向	权	实验考虑权重	标签集	实验重新切片
合成数据集	5000	20000	25	3	5	无	无	否	是	否
DGraphFin数据集	3700500	4300999	17	821	4	有	无	否	是	是

巍巍天山雪，泱泱红湖浪
高山人仰止，源远水流长

新疆大学
XINJIANG UNIVERSITY

Comparison Algorithm

In the process of providing continuous learning for dynamic community detection, I selected five targeted algorithms for comparison based on the characteristics of my proposed algorithm:

1. KMeans: It is a very popular clustering algorithm used for community detection, and introducing temporal characteristics can be used for dynamic community detection. This algorithm groups data in the form of k clusters and optimizes the objective function through iteration to minimize the sum of squared distances between each data point and the centroid of the cluster it belongs to. What I commonly use here is to use the node centrality input method in the graph for processing.

2. DyLouvain: It is a dynamic community detection method based on the Louvain algorithm for dynamic network community detection, which introduces temporal characteristics to Louvain to process dynamic community snapshots. The Louvain algorithm is a community detection method that recursively merges communities into a single node and performs modular clustering on compressed graphs. This algorithm starts from a weighted network of N nodes, and in the first stage, assigns a different community to each node of the network. Then for each node, it considers neighbors and evaluates modular gains by removing specific nodes from the current community and placing them in the neighboring community. If the gain is positive and maximized, the node will be placed in the neighboring community. Repeat this process until the community to which all nodes belong no longer changes.

Comparison Algorithm

3. LPA: It is a community testing method. The main idea of this method is to iteratively update the labels of each node in the network to discover the community structure in the network. Specifically, the LPA algorithm first sets a unique label for each node, and then iteratively updates each node in sequence. In each iteration, for each node, the label with the highest number of labels is selected by counting the labels of its neighboring nodes to update that node. If the maximum number of labels is greater than 1, randomly select one of these labels to update the node until the algorithm converges.

4. Decs: It is a community detection method based on a dynamic network community evolution model. This method discovers the structure of each club by tracking the changes and interactions among its members. It first performs static community detection at each time point, and then matches and evolves the community based on the changes in its members. The DECS algorithm can effectively discover community structures in dynamic networks and analyze the evolution process and results of communities. It belongs to one of the dynamic community detection methods and is a dynamic community detection algorithm based on the community evolution model.

Comparison Algorithm

5. Chimera: It is a method of detecting and predicting community structure using links and content in the network. It is an efficient algorithm based on shared matrix factorization, which uses links and content that change over time; The unified nature of embedding allows the use of any traditional clustering algorithm on the corresponding representation.

I chose these 5 methods as comparative algorithms to comprehensively verify the effectiveness of multidimensional feature information in dynamic community detection.

Evaluation Indicators

1. Modularity: Modularity is a method of measuring the strength of network community structure, first proposed by Mark Newman. It is a standard for measuring the quality of network template (also known as node set or community) partitioning, and the closer its value is to 1, the stronger the community structure partitioned by the network, that is, the better the partitioning quality. Therefore, the optimal network community partitioning can be achieved by maximizing modularity.

2. Jaccard: The Jaccard evaluation index, also known as Jaccard similarity coefficient or Jaccard similarity, is a statistical indicator used to compare the similarity between two sample sets. The value range of the Jaccard indicator is between 0 and 1. The closer the value is to 1, the more similar the two sets are, and the closer the value is to 0, the less similar the two sets are. In fields such as machine learning and information retrieval, the Jaccard metric is often used to evaluate the similarity between two sets, such as evaluating the performance of tasks such as text classification and clustering.

3. Time Consumption: It is the time it takes for the algorithm to process each network snapshot.

Note: For the three evaluation indicators mentioned above, Q (Modularity), J (Jaccard), and T (Time Consumption) are used as substitutes in the following text, and EI in the table represents the evaluation indicators.

Experimental Result

Table 2: Experimental results of Synthetic Dataset

ALG	EI	1	2	3	ALG	EI	1	2	3	ALG	EI	1	2	3
KMeans	Q	0.0002	0.0003	0.0004	LPA	Q	0	0	0	Chimera	Q	0.4425	0.4427	0.4427
	J	0.1083	0.1098	0.1073		J	0.04	0.04	0.04		J	0.9791	0.9093	0.9621
	T	0.30	0.29	0.27		T	0.31	0.36	0.32		T	0.253	0.283	0.282
DyLouvain	Q	0.4452	0.4464	0.4471	Decs	Q	0.4452	0.4464	0.4471	Dcdmf	Q	0.4415	0.4428	0.4437
	J	0.4	0.0	0.2		J	0	0.2	0.2		J	0.9932	0.9932	0.9936
	T	2.77	2.70	3.25		T	37.05	36.39	40.67		T	0.20	0.15	0.18

Table 3: Experimental results of DGraphFin Dataset

ALG	EI	1	2	3	ALG	EI	1	2	3	ALG	EI	1	2	3
k-means	Q	-0.017	-0.058	-0.076	LPA	Q	0.88	0.83	0.87	Chimera	Q	0.0408	0.0486	0.0342
	J	0.074	0.104	0.060		J	0	0	0		J	0.152	0.153	0.147
	T	44.30	42.40	36.80		T	51.80	64.40	46.90		T	6.20	7.31	7.02
DyLouvain	Q	0.99	0.99	0.99	Decs	Q	0.003	0.0038	0.0034	Dcdmf	Q	0.00038	0.00028	0.00068
	J	0	0	0		J	0.065	0.099	0.098		J	0.265	0.268	0.283
	T	417.40	603.40	457.10		T	784.47	1045.44	978.79		T	4.11	4.46	4.48

巍巍天山雪，泱泱红湖浪
高山人仰止，源远水流长

新疆大学
XINJIANG UNIVERSITY

End

Welcome experts to criticize and correct us

安检入口