
Unsupervised Constrained Community Detection via Self-Expressive Graph Neural Network

Sambaran Bandyopadhyay^{*1}

Vishal Peter^{†2}

¹Amazon Science, Bangalore, India

²Indian Institute of Science, Bangalore, India

Abstract

Graph neural networks (GNNs) are able to achieve promising performance on multiple graph downstream tasks such as node classification and link prediction. Comparatively lesser work has been done to design GNNs which can operate directly for community detection on graphs. Traditionally, GNNs are trained on a semi-supervised or self-supervised loss function and then clustering algorithms are applied to detect communities. However, such decoupled approaches are inherently sub-optimal. Designing an unsupervised loss function to train a GNN and extract communities in an integrated manner is a fundamental challenge. To tackle this problem, we combine the principle of self-expressiveness with the framework of self-supervised graph neural network for unsupervised community detection for the first time in literature. Our solution is trained in an end-to-end fashion and achieves state-of-the-art community detection performance on multiple publicly available datasets.

1 INTRODUCTION

Graphs or networks are ubiquitous in our daily life. Graph representation learning [Perozzi et al., 2014, Hamilton et al., 2017b] is the task of mapping different components of a graph (such as nodes, edges or the entire graph) to a vector space to facilitate downstream graph mining tasks. Among various types of graph representation techniques, graph neural networks (GNNs) [Wu et al., 2020] have received significant attention as they are able to apply neural networks directly on the graph structure. Most of the GNNs can be represented in the form of a message passing network, where

each node updates its vector representation by aggregating messages from neighboring nodes with its own [Gilmer et al., 2017, Hamilton et al., 2017a]. GNNs are traditionally trained in a semi-supervised way [Kipf and Welling, 2017] on a node classification loss when a subset of node labels are available. More recently, unsupervised and self-supervised graph neural networks have been proposed where a reconstruction loss [Kipf and Welling, 2016, Bandyopadhyay et al., 2020] or noise contrastive loss [Veličković et al., 2019, Zhu et al., 2020] is used to train the networks.

Community detection is one of the most important tasks for network analysis and has been studied for decades in classical network analysis community [Fortunato and Hric, 2016, Xie et al., 2013]. However, compared to other tasks such as node classification [Kipf and Welling, 2017, Veličković et al., 2019] and link prediction [Kipf and Welling, 2016, Zhang and Chen, 2018], community detection has not been explored much in the framework of graph neural networks. Being inherently unsupervised in nature, it is challenging to train GNNs for community detection directly. Traditionally, methods have been proposed where a graph representation learning algorithm is trained on a generic unsupervised loss and then a clustering algorithm is applied as a post-processing step to discover communities [Perozzi et al., 2014, Bandyopadhyay et al., 2019]. Such approaches are sub-optimal in nature as the node representation learning module and the clustering algorithm work independently. More recently, there have been efforts to train graph neural networks directly for community detection in graphs [Bo et al., 2020, Zhang et al., 2020] (Section 2).

In contrast to a fully unsupervised approach, a graph neural architecture is proposed in [Chen et al., 2019] for a supervised version of community detection. In classical machine learning, constraint clustering has been shown to be very efficient where must-link or no-link constraints are given as input [Wagstaff et al., 2001]. But, obtaining direct ground truth community labels or such pair-wise constraints is expensive for real-world networks. In this paper, we aim to derive such constraints in an unsupervised way, by using

^{*}Sambaran was affiliated to Indian Institute of Science and IBM Research when the work was conducted.

[†]Both the authors contributed equally.

the principle of self-expressiveness of data [Ji et al., 2014]. This allows to express each data point by a linear combination of other data points which potentially lie in the same subspace. The principle of self-expressiveness has been successfully applied in computer vision and image processing for object detection and segregation [Zhang et al., 2019a, Li and Vidal, 2015]. However, inherent computational demand to build pair-wise similarity matrix and subsequent use of spectral clustering makes it infeasible to directly apply the principle of self-expressiveness and subspace clustering in domains like graphs where number of nodes can be very large [Elhamifar and Vidal, 2009, Ji et al., 2014, 2017]. We have taken a different approach in this paper to address these computational challenges. Our solution uses a self-supervised GNN and generate node communities from the embeddings obtained. To guide the generated communities, we use the principle of self-expressiveness on randomly sampled batches of nodes to generate a set of soft must-link and no-link constraints.

Following are the novel contributions made in this paper:

- We propose a novel community detection algorithm, referred as *SEComm* (Self-Expressive Community detection in graph). To the best of our knowledge, we are the first in literature to combine the principle of self-expressiveness with the framework of self-supervised graph neural network for unsupervised community detection. Our solution is able to use both link structure and the node attributes of a graph to detect node communities.
- To address the computational issues, our solution uses the principle of self-expressiveness to generate a set of *soft must-link or no-link constraints* on a subset of nodes divided into batches. In contrast to existing literature on self-expressiveness (which typically applies spectral clustering as a post-processing step), our solution is trained in an end-to-end fashion.
- To show the merit of the proposed algorithm, we conduct experiments with multiple publicly available graph datasets and compare the results with a diverse set of algorithms. *SEComm* is able to improve the state-of-the-art performance of unsupervised community detection with a significant margin in almost all the real-world datasets we used. Model ablation study and sensitivity analysis give further insights of the algorithm. Source code of *SEComm* is available at <https://github.com/viz27/SEComm>.

2 RELATED WORK

We have presented the related work into three categories.

Graph Neural Networks: Graph neural networks have gained significant attention in last few years with their success on a diverse set of applications [Wu et al., 2020,

Desai et al., 2021]. Typically, GNNs are trained on node-classification, link prediction and graph reconstruction losses [Kipf and Welling, 2017, Hamilton et al., 2017a]. Recently, self-supervised learning has been able to achieve performance close to supervised learning for multiple downstream tasks [Belghazi et al., 2018, Hjelm et al., 2019]. Extending the concept of information maximization, DGI [Veličković et al., 2019] and GRACE [Zhu et al., 2020] have been proposed where information between different graph entities (graph-level to node-level, corrupted versions of a graph etc.) are maximized. However, none of the GNNs above handles community detection in their respective objectives.

Principle of Self-Expressiveness: The concept of self-expressiveness was proposed to cluster data drawn from multiple low dimensional linear or affine subspaces embedded in a high dimensional space [Elhamifar and Vidal, 2009]. Given enough samples, each data point in a union of subspaces can always be written as a linear or affine combination of all other points [Elhamifar and Vidal, 2009, Ji et al., 2014]. Subspace clustering exploits this to build a similarity matrix, from which the segmentation of the data can be easily obtained using spectral clustering [Lu et al., 2012, Ji et al., 2014]. Recently, a deep learning based subspace clustering method has been proposed where an encoder is used to map data to some embedding space before building the pair-wise similarity matrix and applying spectral clustering [Ji et al., 2017]. However, inherent computational demand to build pair-wise similarity matrix and subsequent use of spectral clustering makes it infeasible to directly apply the principle of self-expressiveness in domains like graphs where the number of nodes can be very large.

Community Detection with GNNs: As explained in Section 1, a disjoint approach of applying clustering on node embeddings obtained by some representation algorithm is inherently sub-optimal in nature [Bandyopadhyay et al., 2020]. In [Zhang et al., 2019b], the authors have used an adaptive graph convolution method that performs high-order graph convolution to obtain smooth node embeddings that capture global cluster structure. The node embeddings obtained are subsequently used to detect communities using spectral clustering. In [Sun et al., 2019], a probabilistic generative model is proposed to learn community membership and node representation collaboratively. More recently, researchers have tried to propose GNN algorithms that can operate directly for community detection in a graph [Bo et al., 2020]. In [Zhang et al., 2020], authors propose to derive node community membership in the hidden layer of an encoder and introduced a community-centric dual decoder to reconstruct network structures and node attributes in an unsupervised fashion. Our work is towards this direction of obtaining node communities directly in the framework of graph neural networks.

3 PROBLEM FORMULATION

Let us denote an input graph by $G = (V, E, X)$, where $V = \{1, 2, \dots, N\}$ is the set of nodes and $E \subseteq V \times V$ is the set of edges. We assume that each node has some attribute values present in a vector $x_i \in \mathbb{R}^F$ and $X = [x_1 \ x_2 \ \dots \ x_N]^T \in \mathbb{R}^{N \times F}$ is the node attribute matrix of the graph. The goal of our work is to learn a function $f : V \mapsto [K]$, where $[K] = \{1, 2, \dots, K\}$ is the set of community (or cluster) indices, to map each node to a community by exploiting the link structure and node attributes of the graph. We want to achieve this without having any ground truth community information of a node. Intuitively, nodes which are closely connected in the graph or have similar attributes, should be members of the same community. Important notations in the paper are summarized in Table 1.

Notations	Explanations
$G = (V, E, X)$	Input graph
$i, j \in \{1, 2, \dots, N\} = V$	Indices over nodes
$X \in \mathbb{R}^{N \times F}$	Node feature matrix
$Z \in \mathbb{R}^{N \times F'}$	Node embedding matrix
M, P	Batch size and number of batches sampled
$S_{ij} \in [0, 1]$	Similarity between nodes i and j
$\mathcal{S} \subseteq V \times V$	node pairs whose similarities are computed
$\mathcal{S}_{ext} \subseteq \mathcal{S}$	After filtering out with θ_{low} and θ_{high}
$C_i \in \mathbb{R}^K$	Community membership vector for i th node
$C \in \mathbb{R}^{N \times K}$	Output node community membership matrix
W_{SS}, W_{MLP}	Parameters of SS-GNN and MLP modules

Table 1: Different notations used in the paper

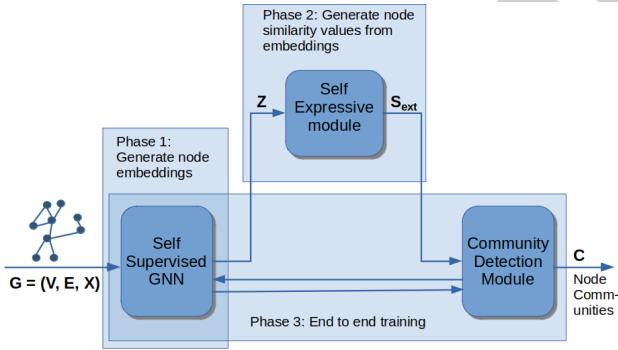


Figure 1: Training phases of SEComm

4 OUR SOLUTION: SECOMM

There are multiple steps in our proposed solution SEComm as shown in Figure 1. We discuss each of them.

4.1 SELF-SUPERVISED NODE EMBEDDING

The first step of SEComm is to learn node representation in an unsupervised way. Self-supervised learning [Hjelm

et al., 2019] has been used recently for obtaining both node embeddings [Veličković et al., 2019, Zhu et al., 2020] and graph-level embeddings [Sun et al., 2020]. Potentially, any self-supervised differentiable approach to obtain node representation can be integrated with our solution. In the following, we have adopted the principle of mutual information maximization between two corrupted versions of the given graph, motivated from [Veličković et al., 2019, Zhu et al., 2020], which is then used to formulate the final objective of SEComm in Section 4.3.

Given the input graph $G = (V, E, X)$, two graph views G_1 and G_2 are generated from it by employing a corruption function. The corruption function randomly removes a small portion of edges from the input graph and also randomly masks a fraction of dimensions with zeros in node features. The vertex sets of G_1 and G_2 remain the same. These views are used for contrastive learning at both graph topology and node feature levels. We use a GCN encoder to generate node embeddings for both G_1 and G_2 . For a graph G , GCN encoder derives node representations as follows:

$$Z = f(X, A) = \text{ReLU}(\hat{A} \text{ReLU}(\hat{A}XW^{(0)}))W^{(1)} \quad (1)$$

where each row of $Z \in \mathbb{R}^{|V| \times F'}$ contains the corresponding node representation. A is the adjacency matrix of the graph G . We compute $\tilde{A} = A + I$, where $I \in \mathbb{R}^{|V| \times |V|}$ is the identity matrix and the degree diagonal matrix \tilde{D}_{ii} with $\tilde{D}_{ii} = \sum_{j \in V} \tilde{A}_{ij}$, $\forall i \in V$. We set $\hat{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$. $W^{(0)}$ and $W^{(1)}$ are the trainable parameter matrices of GCN. Let us use Z_1 and Z_2 to denote the node embedding matrices for the two views G_1 and G_2 obtained from the GCN encoder (parameters shared).

Next, the following noise contrastive objective (via a discriminator) is used. For any node $i \in V$, let us denote the corresponding nodes in G_1 and G_2 as $G_1(i)$ and $G_2(i)$ respectively. For each $i \in V$, the pair $(G_1(i), G_2(i))$ is considered as a positive example. Negative examples are sampled from both the views for each node $i \in V$. More formally, we randomly select a set of nodes $V_{-i} = \{j \in V \mid j \neq i\}$ such that $|V_{-i}| = N_-$ (number of negative samples), $\forall i \in V$. Both $(G_1(i), G_1(j))$ and $(G_1(i), G_2(j))$ are considered as negative examples. Then, the following objective function is minimized:

$$\min_{W_{SS}} \mathcal{L}_{SS} = \sum_{i \in V} \left[-\frac{\cos(Z_{1i}, Z_{2i})}{\tau} + \log \left(\sum_{j \in V_{-i}} e^{\frac{\cos(Z_{1i}, Z_{1j})}{\tau}} + e^{\frac{\cos(Z_{1i}, Z_{2j})}{\tau}} \right) \right] \quad (2)$$

where Z_{1i} and Z_{2i} denotes the i th row of Z_1 and Z_2 respectively, $\cos()$ is the cosine similarity between the two embeddings and τ is a temperature parameter. This essentially maximizes the agreement between the embeddings

of i th node in two views. Both link structure and node attributes of the graph are considered because of the use of GCN encoder in Equation 1. Node embedding matrix Z of the original input graph G can be obtained by Equation 1, where the purpose of G_1 and G_2 were just to enable the training of the self-supervised loss in Equation 2.

4.2 LEARNING NODE SIMILARITIES THROUGH SELF-EXPRESSIVENESS

Vector node representations obtained in Section 4.1 are quite generic. There is no guarantee that similarities between the nodes captured through such embeddings are suitable to discover communities in the graph. In contrast to node classification or other supervised tasks, lack of any ground truth in training for community detection makes the problem highly non-trivial. To tackle this, we use the principle of self-expressiveness [Elhamifar and Vidal, 2009, Ji et al., 2017] which aims to approximate a data point by a sparse linear sum of a subset of other points, which stays in the same subspace. This is more prevalent when the graph has large number of nodes and embedding dimension is also reasonably high [Elhamifar and Vidal, 2009]. Based on the contribution of a point to reconstruct some other point, it is possible to learn a pair-wise similarity matrix using this principle. Such a pair-wise similarity can guide the generation of communities from the node embeddings obtained by the self-supervised layer. However, computation of pair-wise similarity matrix for a graph can be too expensive as it needs $O(N^2)$ storage and processing. Hence, we propose a batch-wise learning procedure, as discussed later.

Given the node embedding matrix $Z \in \mathbb{R}^{N \times F'}$ obtained from Section 4.1, we want to derive a node similarity matrix $S \in \mathbb{R}^{N \times N}$ using the principle of self-expressiveness. For each node $i \in V$, we first try to express Z_i (i th row of Z) by a linear sum of few other node embeddings Z_j , $j \neq i$. So, $Z_i = \sum_{j \in V} q_{ij} Z_j$, where q_{ij} is the (i, j) th element of a coefficient matrix $Q \in \mathbb{R}^{N \times N}$ and we enforce $Q_{ii} = 0$ to avoid the trivial solution of Q being assigned to a identity matrix. We need to learn this coefficient matrix Q which will be used to generate similarity matrix S . It can be shown [Ji et al., 2014] under the assumption of subspace independence that, by minimizing certain norms of Q , it is possible to have a block-diagonal structure (up to a permutation) of Q . In that case, each block in Q would contain nodes which belong to the same subspace. This can be posed as the following optimization problem.

$$\min_Q \|Q\|_p \text{ such that, } Z = QZ; \text{diag}(Q) = 0 \quad (3)$$

where $\|Q\|_p$ is p th matrix norm of Q and $\text{diag}(Q)$ denotes the diagonal entries of Q . Based on the choice in some existing literature [Lu et al., 2012, Ji et al., 2014], we use square Frobenius matrix norm for our implementation. However,

exact reconstruction of Z using the this principle may not be possible. So, we relax the hard constraint $Z = QZ$ with square Frobenius norm of $(Z - QZ)$ (soft constraint). This gives us the following objective function.

$$\min_Q \mathcal{L}_{SE} = \|Z - QZ\|_F^2 + \lambda_1 \|Q\|_F^2 \text{ such that, } \text{diag}(Q) = 0 \quad (4)$$

where λ_1 is a weight parameter of this optimization.

In principle, while a pairwise similarity matrix S can be constructed trivially as $Q+Q^T$, many heuristics have been proposed to improve the clustering performance of S (when using methods such as spectral clustering directly on S). We follow the heuristics proposed in [Ji et al., 2014] to construct the node similarity matrix S as:

1. $Q' = \frac{1}{2}(Q + Q^T)$
2. Compute the r rank SVD of Q' , ie. $Q' = U\Sigma V^T$, where $r = dK + 1$, K is the number of communities and d is the maximal intrinsic dimension of subspaces which is set to 4 in all our experiments.
3. Compute $L = U\Sigma^{\frac{1}{2}}$ and normalize each row of L to have unit norm.
4. Set negative values in L to zero to obtain L' .
5. Construct similarity matrix S as $S = (L' + L'^T)/\|L\|_\infty$, so that $s_{ij} \in [0, 1]$.

As mentioned before, an inherent difficulty to compute the pair-wise similarity matrix is the computation and storage of $N \times N$ dimensional matrix S . So, instead of computing this matrix for all pairs of nodes, we use batch-wise learning. We sample batches of randomly selected nodes with batch size M , where $M \leq N$. We train the loss in Equation 4 for each batch. The required computation in each batch is $O(M^3)$ (for solving Equation 4 and the subsequent use of SVD decomposition) which is much lesser than $O(N^3)$ for a significantly smaller M . However, the problem with this approach is that one would not get complete similarity matrix for the graph. It only computes s_{ij} if nodes i and j belong to a same batch. Let us denote \mathfrak{S} to be the set of node pairs for which the similarity is computed in the batch-wise learning. Clearly, $\mathfrak{S} \subsetneq V \times V$ and $|\mathfrak{S}| \ll N^2$ (when $M < N$). This makes it difficult to use with spectral clustering, as most of the subspace clustering algorithms do [Ji et al., 2017]. But as explained next, our overall solution does not need all the node-pair similarities. Rather, it filters the existing similarities computed with the batch-wise solution using a simple trick explained next.

4.3 CONSTRAINED NODE COMMUNITY DETECTION

Instead of applying expensive spectral clustering on the complete matrix S as a post processing step to find node clusters, we use a neural network based solution which is

significantly more scalable. We use a fully-connected multi-layer perceptron (MLP) with the set of trainable parameters W_{MLP} to map each node embedding to its corresponding soft community membership vector as follows.

$$C_i = \text{Softmax}(\text{MLP}(Z_i)) \in \mathbb{R}^K \quad (5)$$

where the MLP maps each $Z_i \in \mathbb{R}^{F'}$ to a K dimensional vector, K is the number of communities. We assume to know K beforehand. The softmax layer converts the K dimensional vector to a probability distribution such that c_{ik} (k th element of C_i) denotes the probability that i th node belongs to k th community, $\forall k = [K]$. Equation 5 ensures that nodes having similar embeddings will be mapped to similar positions in the $(K - 1)$ dimensional probability simplex. However, relying completely on embeddings to detect communities is not desirable since the embeddings are generated with generic objectives. So, they may not be optimal to generate node communities. Hence, we use the information learned in node-pair similarities in Section 4.2 to guide both the detection of node communities by training the parameters of MLP in Equation 5 and updating node embeddings.

Let us form the community membership matrix $C = [C_1, \dots, C_N]^T \in \mathbb{R}^{N \times K}$. If the complete node similarity matrix S is available, one may try to minimize the following objective.

$$\min_{W_{SS}, W_{MLP}} \|CC^T - S\|_F^2 \quad (6)$$

There are multiple drawbacks present in the objective function above. First, it needs us to compute the complete node similarity matrix S in Section 4.2 which prevents the batch-learning mechanism explained before. Next, the computation involved is $O(N^2)$ in Equation 6. Further, there is another issue if one wants to use all pair-wise node similarities in S to guide the community detection. Due to noise in the dataset, many of the pair-wise similarities may not reflect the actual similarities between the nodes. The similarity values which are around 0.5 neither express a strong similarity nor a strong dissimilarity between a node pair. So they are less informative compared to the similarity values which are close to 0 or 1. But they can still influence the parameters of the neural network because of Equation 4.

Hence, instead of considering all the pair-wise similarity values, we only consider the ones in \mathfrak{S} computed over the batches as discussed in Section 4.2. Further, we have observed experimentally in Section 5.6 that for a larger dataset, it is okay even if some nodes are not part of any of the batches selected randomly. Thus, the number of batches can be significantly smaller than $\frac{N}{M}$ for a larger dataset. We also introduce two thresholds θ_{low} and θ_{high} to use only those node-pair similarities which are extreme in their values, thus more informative in nature. We set $0 < \theta_{low} \leq \theta_{high} < 1$. We also set $\theta_{high} = 1 - \theta_{low}$, as this choice works well in the experiments and also reduces the number of hyper-parameters. Let us introduce the set $\mathfrak{S}_{ext} \subseteq \mathfrak{S}$ as follows.

$$\mathfrak{S}_{ext} = \left\{ (i, j) \in \mathfrak{S} \mid S_{ij} \leq \theta_{low} \text{ or } S_{ij} \geq \theta_{high} \right\} \quad (7)$$

Here, a node pair (i, j) in \mathfrak{S}_{ext} should be roughly constrained to be in the same cluster when S_{ij} value is very high or in different clusters when S_{ij} is very low. Thus, we derive a set of soft version of must-link and no-link constraints in an unsupervised way to guide the formation of communities. With these, we formulate the following optimization to detect the communities:

$$\min_{W_{SS}, W_{MLP}} \sum_{(i,j) \in \mathfrak{S}_{ext}} (C_i^T C_j - S_{ij})^2 \quad (8)$$

By considering only the node pairs in \mathfrak{S}_{ext} , we are able to ignore the pairs which are neither too similar nor too dissimilar, to contribute to the learning of community memberships. As C_i is a probability distribution over all the K communities for a node i , we want to avoid trivial community formations where each node is assigned to all the communities with roughly uniform probabilities, or all the nodes are assigned to a single community [Bianchi et al., 2020]. So, we update the main objective in Equation 8 as:

$$\min_{W_{SS}, W_{MLP}} \mathcal{L}_{Com} = \sum_{(i,j) \in \mathfrak{S}_{ext}} (C_i^T C_j - S_{ij})^2 + \lambda_2 \left\| \frac{C^T C}{\|C^T C\|_F^2} - \frac{I_K}{K} \right\|_F^2 \quad (9)$$

The second component in the equation above ensures that communities are close to orthogonal and they are balanced in size. Please note that due to the use of neural network to generate community membership for each node in Equation 5, the optimization in Equation 9 is not a discrete optimization. Rather, we solve it with respect to the parameters W_{SS} of the self-supervised layer (Eq. 2) and W_{MLP} of the MLP (Eq. 5). The total loss to train the node embeddings and community detection can be written as a weighted sum of self-supervised loss and community detection loss, which is shown below:

$$\min_{W_{SS}, W_{MLP}} \mathcal{L}_{total} = \alpha \mathcal{L}_{SS} + \mathcal{L}_{Com} \quad (10)$$

where α is a weight factor of the optimization. The node-pair similarity values are obtained by solving the batch-learning technique in Section 4.2. The overall algorithm SEComm proceeds in an iterative way by solving the self-expressive layer for each batch and then updating the parameters of the neural network by minimizing Equation 10. The pseudo code of SEComm is presented in 1.

4.4 TRAINING AND ANALYSIS OF SECOMM

We use ADAM with default parameterization to solve the optimization formulations in Equations 4 and 10. For the

Algorithm 1 SEComm - Self-Expressive Community Detection

Input: The graph $G = (V, E, X)$, $|V| = N$, K : Number of communities in the graph, M : Batch size for the self-expressive layer, P : Number of batches used for training self-expressive layer, Thresholds θ_{low} and θ_{high} .

Output: Community membership vector $C_i \in \mathbb{R}^K$ for each node $i \in V$.

- 1: Initialize the parameters of the self-supervised GNN and clustering MLP (in Eq. 5).
- 2: pre-training step: Obtain node embeddings $Z \in \mathbb{R}^{N \times F'}$ by training the self-supervised GNN.
- 3: Initialize \mathfrak{S} as empty.
- 4: **for** $batch \in \{1, 2, \dots, P\}$ **do**
- 5: Sample a batch of M nodes from V
- 6: Learn the pair-wise node similarity matrix S_m for the selected nodes by optimizing Eq. 4.
- 7: Add all the node-pairs from the batch to \mathfrak{S} .
- 8: **end for**
- 9: Construct \mathfrak{S}_{ext} according to Eq. 7
- 10: **for** $iter \in \{1, 2, \dots, T\}$ **do**
- 11: Generate node embedding matrix $Z \in \mathbb{R}^{N \times F'}$ using the self-supervised GNN.
- 12: Generate cluster membership vector $C_i \in \mathbb{R}^K$ for each node $i \in V$.
- 13: Update the parameters of the GNN and clustering MLP by optimizing Eq. 10
- 14: **end for**

self-expressive loss in Equation 4, we train until the loss saturates. For the total loss in Equation 10, we particularly focus

on the saturation of the regularization $\left\| \frac{C^T C}{\|C^T C\|_F^2} - \frac{I_K}{K} \right\|_F^2$.

Experimentally, using the convergence of this component explicitly as a stopping criteria for SEComm gives slightly better result for all the datasets, than checking the total convergence. But as explained in Section 5.3, different components of the loss function have similar contributions. Hence they saturate almost in the same time for most of the cases. This can also be observed in Section 5.5.

Time Complexity: Time complexity of the self-supervised GNN in Section 4.1 is $O(|E| + NFF'N_-)$, where N_- is the number of negative samples used. The self-expressive layer takes another $O(PM^3)$ time, where P is the number of batches sampled, and M is the size of each batch. Finally, generating community membership takes $O(NK^2)$ time because of solving the loss in Equation 9. Thus, the overall run time of each iteration of SEComm is linearly dependent on the number of nodes and number of edges in the graph.

5 EXPERIMENTAL EVALUATION

This section presents the details of the experiments that we conducted and the analysis of the results.

Dataset	#Nodes	#Edges	#Features	#Labels
Cora	2,708	5,429	1,433	7
Citeseer	3,327	4,732	3,703	6
Pubmed	19,717	44,338	500	3
Wiki	2,405	17,981	4,973	17
Physics	34,493	247,962	8,415	5

Table 2: Summary of the datasets used

5.1 DATASETS USED

To show the merit of SEComm, we conduct experiments on 5 publicly available graph datasets [Kipf and Welling, 2017, Zhang et al., 2019b]. Different statistics of the datasets are summarized in Table 2. Cora, Citeseer and Pubmed are citation datasets where nodes correspond to papers and are connected by an edge if one cites the other. Wiki is a collection of webpages where nodes are webpages and are connected if one links to other. Physics is a co-authorship network where nodes are authors, that are connected by an edge if they have co-authored a paper [Shchur et al., 2018]. Each of these datasets have attribute vector associated with each node. They also have ground truth community membership of each node, which we use to evaluate the performance of our proposed and baseline algorithms.

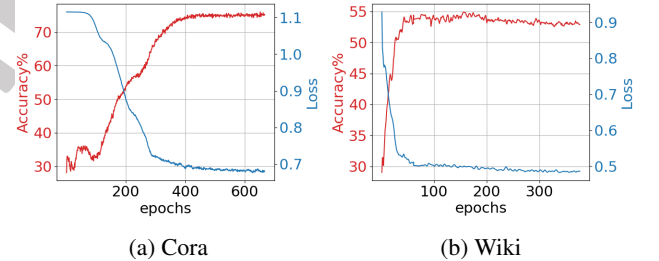


Figure 2: Loss vs Accuracy progression during training

5.2 BASELINE ALGORITHMS

We use a diverse set of baselines to compare the performance of SEComm. We divide them into the following categories.

Using only Node Features: As each node is associated with some attribute vectors, we use **k-means** and spectral clustering (**Spectral-f**) algorithms directly on the node attributes to cluster nodes into different communities. Naturally, these approaches ignore the graph structure completely.

Using only Graph Structure: We also use spectral clustering (**Spectral-g**) on the graph structure. Here we consider adjacency matrix of a graph as the similarity matrix between the nodes. We also use popular unsupervised node embedding techniques **DeepWalk** [Perozzi et al., 2014], which is a random walk based technique and **DNGR** [Cao et al., 2016], which is an auto-encoder based technique.

Methods	Input	Cora			CiteSeer			Pubmed			Wiki		
		Acc%	NMI%	F1%	Acc%	NMI%	F1%	Acc%	NMI%	F1%	Acc%	NMI%	F1%
k-means	Feature	34.65	16.73	25.42	38.49	17.02	30.47	57.32	29.12	57.35	33.37	30.20	24.51
Spectral-f	Feature	36.26	15.09	25.64	46.23	21.19	33.70	59.91	32.55	58.61	41.28	43.99	25.20
Spectral-g	Graph	34.19	19.49	30.17	25.91	11.84	29.48	39.74	3.46	51.97	23.58	19.28	17.21
DeepWalk	Graph	46.74	31.75	38.06	36.15	9.66	26.70	61.86	16.71	47.06	38.46	32.38	25.74
DNGR	Graph	49.24	37.29	37.29	32.59	18.02	44.19	45.35	15.38	17.90	37.58	35.85	25.38
GAE	Both	53.25	40.69	41.97	41.26	18.34	29.13	64.08	22.97	49.26	17.33	11.93	15.35
VGAE	Both	55.95	38.45	41.50	44.38	22.71	31.88	65.48	25.09	50.95	28.67	30.28	20.49
MGAE	Both	63.43	45.57	38.01	63.56	39.75	39.49	43.88	8.16	41.98	50.14	47.97	39.20
ARGE	Both	64.00	44.90	61.90	57.30	35.00	54.60	59.12	23.17	58.41	41.40	39.50	38.27
ARVGE	Both	63.80	45.00	62.70	54.40	26.10	52.90	58.22	20.62	23.04	41.55	40.01	37.80
AGC	Both	68.92	53.68	65.61	67.00	41.13	62.48	69.78	31.59	68.72	47.65	45.28	40.36
GUCD	Both	50.5	32.3	NA	54.47	27.43	NA	63.13	26.98	NA	NA	NA	NA
SEComm	Both	75.92	56.04	73.94	69.82	42.53	60.25	74.49	36.50	73.50	53.10	51.38	44.48
Rank	(SEComm)	1	1	1	1	1	2	1	1	1	1	1	1

Table 3: Performance of Community Detection by SEComm and other baseline algorithms

Methods	Acc%	NMI%	F1%	Runtime (sec.)
k-means	44.20	44.46	37.63	28
ARGE	60.67	51.77	62.32	2112
ARVGE	61.28	53.49	65.47	2221
AGC	75.36	58.19	60.72	6931
SEComm	77.93	56.08	76.42	624

Table 4: Community Detection on Physics Dataset

Using both Node Features and Graph Structure: We use a set of unsupervised graph neural network based techniques. GNN based approaches are naturally able to use both link structure and node attribute of the graph. They are: graph autoencoder (**GAE**) and graph variational autoencoder (**VGAE**) [Kipf and Welling, 2016], marginalized graph autoencoder (**MGAE**) [Wang et al., 2017], adversarially regularized graph autoencoder (**ARGE**) and variational graph autoencoder (**ARVGE**) [Pan et al., 2018]. These methods typically learn the node embeddings and use clustering on the embeddings as a post processing step. Finally, we use two recently proposed community detection methods - **AGC** [Zhang et al., 2019b], which uses high-order graph convolution to get node embeddings and detect communities via spectral clustering on the embeddings and **GUCD** [Zhang et al., 2020], which uses an auto-encoder based framework to obtain direct community assignments for every node.

5.3 EXPERIMENTAL SETUP

Our proposed algorithm SEComm generates community membership of each node in a graph in the framework of graph neural networks. As each node has a single ground truth community membership in all the datasets that we use, we consider the index of the maximum value of $C_i \in \mathbb{R}^K$ (from Equation 5) as the community index of the node

generated by SEComm.

There are multiple hyperparameters present in SEComm. For weight factors in optimization such as λ_1 , λ_2 and α , we check the contribution of different components in a loss function at the beginning of the algorithm, and set these parameters to values such that effective contributions of those components become roughly the same. This ensures that the optimization pays similar importance to different components of SEComm. For the temperature parameter τ , we use the same values used in the literature [Zhu et al., 2020]. For threshold parameters θ_{low} ($0 < \theta_{low} \leq 0.5$), we set it to 0.5 for relatively smaller datasets as we do not want to discard any information for them. For Pubmed, we set it to 0.05 as considering more node-pair similarity values adds noise and also increases runtime of SEComm. However on Physics, the training convergence is not smooth when we set θ_{low} to a very small number. So, we set it to 0.3 on this dataset. As mentioned in Section 4.3, we set $\theta_{high} = 1 - \theta_{low}$. We have also conducted sensitivity analysis of SEComm with respect to some of these hyperparameters in Section 5.6.

5.4 RESULTS OF COMMUNITY DETECTION

Tables 3 and 4 show the performance of community detection by different baseline algorithms and SEComm. We use three popularly used metrics to evaluate the quality of community detection. They are clustering accuracy (Acc), normalized mutual information (NMI), and macro F1-score (F1) [Aggarwal and Reddy, 2014, Zhang et al., 2019b]. We use ground truth community information of the nodes only to calculate these quality metrics.

While reporting the performance of baseline algorithms for the first four datasets in Table 3, we have collected the best

results from the available literature [Zhang et al., 2019b, 2020] which adopted the same experimental setup. We mark some entry as ‘NA’ if the result of that algorithm for a dataset is not publicly available. For Physics dataset, the baseline results are not available in the literature. So, we have run and reported results only for better-performing and diverse subset of baselines in Table 4 with adequate hyperparameter tuning. Additionally, we have also reported the runtime on Physics dataset for these algorithms to give more insight about scalability.

We run SEComm 10 times on each dataset and report the average performance. Tables 3 and 4 show that SEComm is able to achieve state-of-the-art (SOTA) performance for all the datasets, and for all the metrics, except on Citeseer-F1% and Physics-NMI% scores, where SEComm is next to AGC. In terms of performance improvement by clustering accuracy, SEComm is able to improve SOTA by 10.1% on Cora, 4.2% on Citeseer, 6.7% on Pubmed, 5.9% on Wiki and 3.4% on Physics. We also check the standard deviation of the performance of SEComm over 10 runs in each dataset. Standard deviation lies in the range of 0.5% - 1% on all the datasets, which shows the robustness of SEComm. Among the baselines, AGC mostly performs better than others. But, AGC is computationally much expensive because of their adaptive strategy for determining k in k -order graph convolution. As shown in Table 4, runtime of AGC on Physics dataset is ~ 16 times more than that of SEComm. As expected, the algorithms which use both graph structure and node attributes perform better than the ones which use only one of those. The consistent performance of SEComm on all the datasets shows the importance of integrating the objective of community detection directly into the framework of self-supervised graph neural network (the loss from these components propagates to each other through backpropagation). Further, use of the principle of self-expressiveness regularizes the communities formed in SEComm to achieve better performance. Run-time of SEComm and its various components on all the datasets is shown in Table 5. Usefulness of the individual components of SEComm are presented in Section 5.7.

Dataset	GNN	Self-Express.	Comm. Module	Total
Wiki	5.4s	34.8s	44.8s	85.1s
Cora	6.6s	39.2s	84.2s	130.1s
CiteSeer	27.8s	58.4s	157.8s	244.1s
Pubmed	239.6s	103.6s	121.9s	465.1s
Physics	106.9s	280.4s	236.4s	623.7s

Table 5: Runtime of SEComm on various datasets

5.5 LOSS AND ACCURACY OF SECOMM

Typically for an unsupervised algorithm, the loss that it minimizes and the metric that is used to evaluate the performance of the algorithm are not necessarily the same. So, it is important to see if reducing the loss over the epochs

actually increases the performance of the algorithm with respect to the quality metric. For SEComm, we plot the loss in Equation 10 and the clustering accuracy that it achieves over different epochs of the algorithm for the datasets Cora and Wiki in Figure 2. One can see that with the decreasing loss, overall clustering accuracy improves with some minor fluctuations. Thus, the unsupervised loss that SEComm minimizes essentially helps to improve the performance of clustering. This is also another reason of consistent performance (improved metric scores with less standard deviation) of SEComm on multiple datasets.

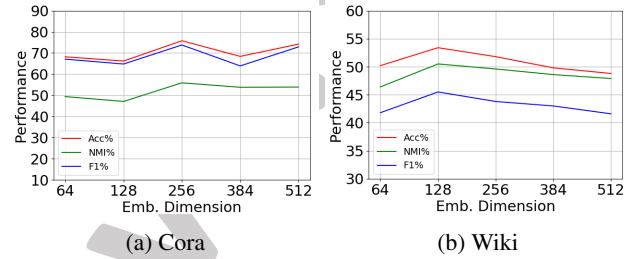


Figure 3: Performance of SEComm with varying embedding dimension

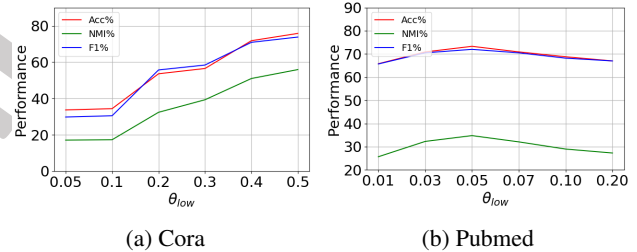


Figure 4: Performance of SEComm with varying threshold θ_{low}

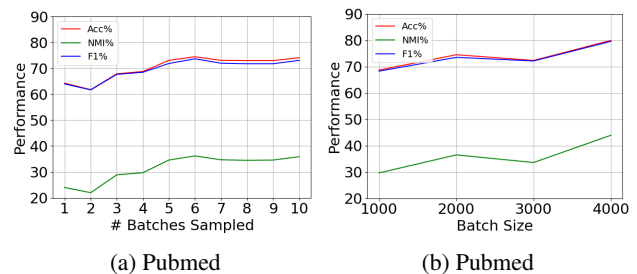


Figure 5: Performance of SEComm with varying number of batches sampled and batch size respectively

5.6 SENSITIVITY TO HYPERPARAMETERS

In this section, we show the sensitivity of SEComm to different hyperparameters. We keep all other hyperparameters fixed while changing only the hyperparameter of interest.

Methods	Cora			CiteSeer			Pubmed			Wiki		
	Acc%	NMI%	F1%	Acc%	NMI%	F1%	Acc%	NMI%	F1%	Acc%	NMI%	F1%
SEComm-GNN	72.85	54.40	68.16	64.95	39.01	60.51	49.63	19.59	40.04	32.14	34.88	28.34
SEComm-Spectral	74.26	55.38	71.75	57.58	36.63	54.99	-	-	-	49.81	51.07	43.95
SEComm-Embeddings	73.85	57.57	65.07	60.29	35.46	54.46	68.30	34.47	67.97	35.67	36.12	29.49
SEComm	75.92	56.04	73.94	69.82	42.53	60.25	74.49	36.50	73.50	53.10	51.38	44.48

Table 6: Model Ablation Study of SEComm

We vary the embedding dimension F' from 64 to 512 for Cora and Wiki and show the performance of community detection in Figure 3. We can see some fluctuation in the performance for Cora. As other hyperparameters are tuned keeping $F' = 256$ for Cora, there are some sudden lows around it. For Wiki, the performance was low at $F' = 64$ as that is not sufficient enough to preserve all the information about the graph in the embeddings. It increases at $F' = 128$. There is a gradual decrease of performance beyond that as the embeddings start holding noisy information when dimension increases more.

We check the performance on Cora, Wiki and Pubmed in Figure 4 with varying θ_{low} (we set $\theta_{high} = 1 - \theta_{low}$). As we decrease θ_{low} , we are filtering out more pair-wise similarities, especially the ones which lies in the mid zone of the range $[0, 1]$. Filtering out such similarity values might lead to less amount of data to regularize the communities learned by SEComm in Equation 9 for a smaller dataset. Thus, the performance on Cora is affected when θ_{low} is very low in Figure 4. But on a larger dataset, filtering out such less informative similarity values (as explained in Section 4.2) can lead to removing noise and help improving the performance. Thus in Pubmed, better performance is observed around $\theta_{low} = 0.05$ (which implies $\theta_{high} = 0.95$). Below which the amount data becomes too less to train the algorithm properly, and above which it was adding noise.

As discussed in Section 6, we do not need to train the self-expressive layer (in Eq. 4) on the complete dataset. In Figure 5a, we vary the number of batches sampled, where each batch contains 2000 nodes. We can see that the performance improved initially and then saturates when number of batches is 5 or more. Thus, optimal performance on Pubmed for community detection can be achieved by using only ~50% (or more) of data points to train the self-expressive layer. In Figure 5b, we change the batch size, keeping number of batches as 6. More is the batch size, more computation resource and time needed. We observe that SEComm is able to achieve reasonably good performance when the batch size is 2000 or more.

5.7 MODEL ABLATION STUDY

In this section, we show the usefulness of different components of SEComm. In particular, we check the community detection performance in the following scenarios.

SEComm-GNN We run k-means on the node embeddings produced by the self-supervised GNN used in SEComm, without running the other modules of SEComm.

SEComm-Spectral We run spectral clustering on the complete similarity matrix S to find node clusters. However, S can be computed only for smaller graphs and hence this experiment cannot be performed on larger datasets like Pubmed and Physics.

SEComm-Embeddings We run k-means on the node embeddings generated after the complete training of SEComm (including the self-expressive and clustering modules)

We compare the results of the above with the community detection output of the complete model of SEComm in Table 6. Again we use three metrics clustering accuracy, NMI and F1 score to evaluate the quality of community detection. Interestingly, we do not see any clear winners between the three model variants. But it is clear from the reported performance numbers that the complete model of SEComm outperforms its variants (except CiteSeer-F1%).

6 DISCUSSION AND CONCLUSION

In this work, we have proposed a novel graph neural network that can directly be used for node community detection in a graph. We use the principle of self-expressiveness to derive a set of soft node-pair constraints to regularize the formation of the communities. To the best of our understanding, this is the first work to integrate a self-expressive layer into a self-supervised GNN. Our approach is highly scalable, without compromising the performance of community detection. SEComm is able to achieve state-of-the-art performance on all the datasets that we used for community detection.

Due to the use of graph neural network to directly generate community memberships of nodes, SEComm can work in an inductive setup. It would be interesting to analyze the performance of SEComm on newly added nodes or even new graphs without retraining.

Acknowledgements

We want to thank Prof. M. Narasimha Murty from CSA, IISc for his feedback on this work.

References

- Charu C Aggarwal and Chandan K Reddy. Data clustering. *Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra*, 2014.
- Sambaran Bandyopadhyay, N Lokesh, and M Narasimha Murty. Outlier aware network embedding for attributed networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 12–19, 2019.
- Sambaran Bandyopadhyay, N Lokesh, Saley Vishal Vivek, and MN Murty. Outlier resistant unsupervised deep architectures for attributed network embedding. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 25–33, 2020.
- Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540, 2018.
- Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral clustering with graph neural networks for graph pooling. In *International Conference on Machine Learning*, pages 874–883. PMLR, 2020.
- Deyu Bo, Xiao Wang, Chuan Shi, Meiqi Zhu, Emiao Lu, and Peng Cui. Structural deep clustering network. In *Proceedings of The Web Conference 2020*, pages 1400–1410, 2020.
- Shaosheng Cao, Wei Lu, and Qionгкаi Xu. Deep neural networks for learning graph representations. In *AAAI*, volume 16, pages 1145–1152, 2016.
- Zhengdao Chen, Lisha Li, and Joan Bruna. Supervised community detection with line graph neural networks. In *International Conference on Learning Representations*, 2019.
- Utkarsh Desai, Sambaran Bandyopadhyay, and Srikanth Tamilselvam. Graph neural network to dilute outliers for refactoring monolith application. In *Proceedings of 35th AAAI Conference on Artificial Intelligence (AAAI'21)*, 2021.
- Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2790–2797. IEEE, 2009.
- Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics reports*, 659:1–44, 2016.
- Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272, 2017.
- Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017a.
- William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017b.
- R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bklr3j0cKX>.
- Pan Ji, Mathieu Salzmann, and Hongdong Li. Efficient dense subspace clustering. In *IEEE Winter Conference on Applications of Computer Vision*, pages 461–468. IEEE, 2014.
- Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian Reid. Deep subspace clustering networks. In *Advances in Neural Information Processing Systems*, pages 24–33, 2017.
- Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.
- Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.
- Chun-Guang Li and Rene Vidal. Structured sparse subspace clustering: A unified optimization framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 277–286, 2015.
- Can-Yi Lu, Hai Min, Zhong-Qiu Zhao, Lin Zhu, De-Shuang Huang, and Shuicheng Yan. Robust and efficient subspace segmentation via least squares regression. In *European conference on computer vision*, pages 347–360. Springer, 2012.
- Shirui Pan, Ruiqi Hu, Guodong Long, Jing Jiang, Lina Yao, and Chengqi Zhang. Adversarially regularized graph autoencoder for graph embedding. *arXiv preprint arXiv:1802.04407*, 2018.
- Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.

- Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.
- Fan-Yun Sun, Meng Qu, Jordan Hoffmann, Chin-Wei Huang, and Jian Tang. vgraph: A generative model for joint community detection and node representation learning. In *Advances in Neural Information Processing Systems*, pages 514–524, 2019.
- Fan-Yun Sun, Jordan Hoffman, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=r1lff2NYvH>.
- Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=rklz9iAcKQ>.
- Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. Constrained k-means clustering with background knowledge. In *Icml*, volume 1, pages 577–584, 2001.
- Chun Wang, Shirui Pan, Guodong Long, Xingquan Zhu, and Jing Jiang. Mgae: Marginalized graph autoencoder for graph clustering. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 889–898, 2017.
- Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- Jierui Xie, 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):1–35, 2013.
- Binbin Zhang, Zhizhi Yu, and Weixiong Zhang. Community-centric graph convolutional network for unsupervised community detection. In *IJCAI*, 2020.
- Junjian Zhang, Chun-Guang Li, Chong You, Xianbiao Qi, Honggang Zhang, Jun Guo, and Zhouchen Lin. Self-supervised convolutional subspace clustering network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5473–5482, 2019a.
- Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems*, pages 5165–5175, 2018.
- Xiaotong Zhang, Han Liu, Qimai Li, and Xiao-Ming Wu. Attributed graph clustering via adaptive graph convolution. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4327–4333. AAAI Press, 2019b.
- Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020.