Engineering Universe for scientific Research and management
ISSN (Online): 2319-3069
Vol. XIV Issue XII
December 2022

A Survey of Various Community Detection Techniques Using Deep Learning

Pradeep Bhawsar[1], Sumit Sharma[2]
Dept. of CSE, Vaishnavi Institute of Technology & Science, Bhopal, India[1, 2]
er.pradeepbhawsar@gmail.com[1], sumit_sharma782022@yahoo.co.in[2]

Abstract

Nowadays, social media sites like Facebook, Instagram, and LinkedIn connect different people, creating an unofficial community map. Differentiating networks in current web-based media charts is a crucial task. The informal organization may be divided into several groups of hubs that have the same behaviors since networks allow us to collect clients that exhibit similar behaviors. This information on people groups can help us focus key information on customers in a specific geographic area and help us make critical decisions. Researchers and experts from various fields have been working to discover a solution to this issue in recent years using a range of different approaches and methodologies. The study project that was suggested has led to an extensive and thorough review of different ways that have been used to address the issue of finding new local areas. By grouping the aforementioned collection of strategies into four different categories—framework factorization, irregular walk, profound learning, and ghostly techniques—this paper offers assessments of the strategies as well as some observations on them. This overview document helps academics get started in the area of informal organizations and identify networks within them. This publication is highly beneficial in light of the substantial amount of work that has been done in the subject over the last couple of years.

***Keywords:* Community Detection, Machine Learning, Deep Learning, Clustering.**

1. Introduction

A network can be defined in a number of ways by computer science researchers, mathematicians, statisticians or physicists. But to visualize any network, the definition from graph theory will be used. According to which there are some users and any interaction between them leads to an edge creation. The interaction between different users/nodes is a result of characteristics such as friendship, relations between family, business relations etc. A network is not the same as a graph because a network can contain much more information or data about users or their interaction than a graph. A network can also be dynamic in nature which means continuously changing the nature of graph structure by addition or deletion of nodes or edges. Communities are found everywhere from simple graph dataset to real world human interactions.

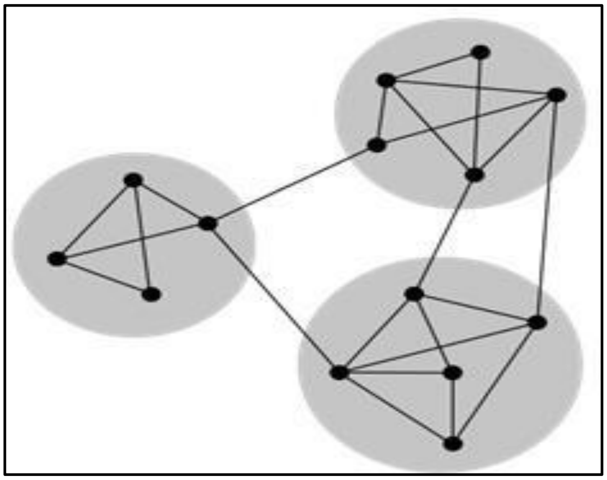

Figure 1 Community structure

Detecting these communities is an important problem as it helps us to gather a lot of information about a cluster. A community can be visualized as a cluster/collection of vertices/users with more interaction between them than other members of the graph structure as shown in Fig1. Users within the same community will exhibit similar behavior and functions. In real life it can be compared with a group of friends or people having the same background or interests. Detecting communities in a graph can also be considered as a graph partitioning problem which comes under the category of NP-hard problems.

Community detection has a variety of applications in a number of different fields. Some of these applications are detecting communities in networks of criminal organizations, analyzing and study of groups which are susceptible to an epidemic disease, also used by companies for dividing market into smaller groups and clusters for advertisement targeting, suggesting products or friends on social media platforms and recommendation system, link prediction and influence maximization. This survey paper summarizes and compares the recent work of researchers and scientists working in the field of community detection in graphs in the last 5-6 years.

This paper is organized as follow: in section 2 we explained related work done previously in the field of community detection, in section 3 we discussed about various techniques used in community detection, in section 4 we conclude this paper with future research directions followed by references used in this paper.

2. Literature review

This section covers the recent progress by researchers and scientists to solve the problem of community detection. After reading research papers and articles of last few years from top conferences and journals, the approaches to solve community detection can be divided into these subcategories - matrix factorization, random walk, deep learning and spectral methods as shown in Fig2. We will go in more detail about work done in each domain in each subsection summarizing various methods and algorithms used for community detection.

A. Deep Learning

The research in the area of deep learning is growing exponentially for the last few years. There is no denying the fact that deep learning has been used to solve various problems in computer science and other fields such as protein folding, capturing black hole images, drug discovery and sports analytics. Neural networks are very powerful tool which can be used to approximate any mathematical function according to universal approximation theorem and has also shown strong representation power. The main contribution of deep learning in solving community detection has been discussed below.

Auto encoders are a very popular and powerful neural network architecture which has very strong representation powers. It consists of two neural networks. One neural network is encoder which encodes our input into lower dimension representations. Other neural networks called decoder tries to reconstruct original data from these lower level representations. It is trained by minimizing the error which is a function of original input and reconstructed output. It is unsupervised technique. The encoder and decoder can be any neural network like simple artificial neural network, convolution neural network or LSTM. This encoding of input into lower representations is the main reason why it has been used in detecting communities in a graph. Yang et al. [1], proposes a deep learning architecture by stacking auto encoders in series. They feed the input modularity matrix into first auto encoder and try to obtain the lower dimensional representation from this encoder by minimizing the reconstructive error loss and feed it into next auto encoder. After a series of similar steps, they finally applied the k-means clustering technique to the encoded output from last auto encoder to detect communities. Dhilber et al. [2] proposes a similar architecture but rather than connecting auto encoders in series and training separately each auto encoder, they stacked them parallel and trained all auto encoders simultaneously.

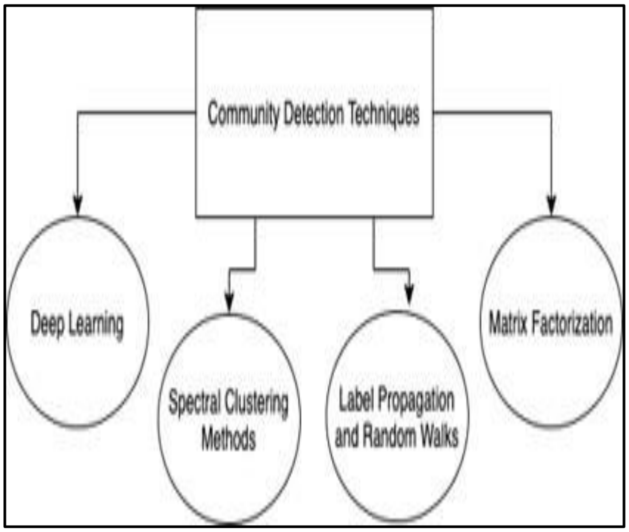
Community Detection Techniques
Deep Learning
Spectral Clustering Methods
Label Propagation and Random Walks
Matrix Factorization

Fig. 2: Community Detection Techniques

Auto encoders are a very popular and powerful neural network architecture which has very strong representation powers. It consists of two neural networks. One neural network is encoder which encodes our input into lower dimension representations. Other neural networks called decoder tries to reconstruct original data from these lower level representations. It is trained by minimizing the error which is a function of original input and reconstructed output. It is unsupervised technique. The encoder and decoder can be any neural network like simple artificial neural network, convolution neural network or LSTM. This encoding of input into lower representations is the main reason why it has been used in detecting communities in a graph. Yang et al. [1], proposes a deep learning architecture by stacking auto encoders in series. They feed the input modularity matrix into first auto encoder and try to obtain the lower dimensional representation from this encoder by minimizing the reconstructive error loss and feed it into next auto encoder. After a series of similar steps, they finally applied the k-means clustering technique to the encoded output from last auto encoder to detect communities. Dhilber et al. [2] proposes a similar architecture but rather than connecting auto encoders in series and training separately each auto encoder, they stacked them parallel and trained all auto encoders simultaneously.

Generative adversarial networks are another powerful deep learning architecture which consists of two neural networks. One is called discriminator and the other one is known as generator. GANs can be considered as a minmax game where generator tries to generate data from real data set and discriminator tries to distinguish between real data and data generated by generator. In this way both tries to minimize their loss by competing with each other trying to fool the other model. GANs and adversarial machine learning has shown successful results in the task of graph embedding and representations. Yuting Jia et al. [3] propose a new architecture CommunityGAN that solves one of the problems arising in traditional methods that is not able to detect overlap- ping communities. Most of the previous techniques believed that one node belongs to only one community that is why they are not able to work on overlapping communities. But CommunityGAN is able to solve this problem. Other than this, the embedding generated by CommunityGAN is able to represent the relation between nodes and communities showing their membership power.

Deep network embedding techniques refers to mapping higher dimensional data which in case is graph into lower level representations in such a way that the data preserves its original structure and information. After obtaining the lower level representation techniques any clustering technique can be applied to obtain communities. Sanjay Kumar et al. [4] used SDNE embedding framework for obtaining the low level representation of input graphs and after that used K Means algorithm optimized with gravitational search algorithm to obtain communities. Sandro Cavallari et al. [5] introduced a new framework where they learn community embedding instead of individual node embedding. They form a cyclic structure of node embedding, community detection and community embedding in which node embedding gives communities, and better communities will help generate better community embedding, and better community embedding will optimize node embedding.

Graph Neural Networks and Graph Convolutional Networks are a new class of neural networks that works on graphs and Non-Euclidean domains. Idea of GCNs is based on convolution neural networks. CNNs usually operate on images by capturing the surrounding information of a pixel of image. On a similar note, here convolution framework of GCNs tries to capture the surrounding information of a node or edge. There has been rapid research development in domain of graph neural networks. Zhengdao Chen et al. [6] propose a new graph neural network called Line graph neural network that solves the community detection problem in a supervised manner and it uses a non backtracking operator which is defined on the edge adjacency list.

B. Label Propagation and Random Walks

Label propagation is a semi-supervised algorithm Zhu and Ghahramani [7] that initially assigns labels to a small subset of the data and then as the algorithm proceeds the labels are propagated to all the unlabeled data points in the space. Random walks is a stochastic process in which an object randomly moves through a mathematical space or structures like a network of connected nodes which can then be used to gain information on the hidden structures (like communities) in the given space.

Krylov Subspace Approximation HE et al. [8] is a technique in which local community is detected by finding a linear sparse coding on the Krylov subspace i.e. the local approximation of spectral subspace. There is a local sampling which uses a seed node to find a comparatively small subgraph Gs in a given graph G. Based on different random short walk diffusion and local community detection by finding sparse relaxed indicator vector lying in local spectral subspace we find the subordinative probability of the corresponding nodes. To get the “Local Spectral subspace” random short walks for probability diffusion from seed set is used instead of eigenvalue decomposition Multiple Community Detection Hollocou et al. [9] technique uses seed nodes to extract communities from the given graph structure. For each seed node a score is calculated to obtain a local community around the seed nodes. The scores are used to define an embedding which maps all the nodes to a vector of appropriate dimension which can be fed to a clustering algorithm DBSCAN M. Ester et al. [10] to obtain K clusters of nodes. Clusters with lower threshold than a specified value are considered the direction to be moving forward, the algorithm moves forward by picking a new seed in each of them. Repeating the steps over multiple iterations till no new seeds formed.

Targeting influential users in a network and then propagating flow of information originating from those nodes with certain probability can help use identify closely grouped nodes. Community detection by simulating information flow Venkatesaramani and Vorobeychik [11] uses this technique to identify community structures in a given graph network. In this technique “Alpha detection” i.e. identifying users that are likely to be the source of information in a network. After these alpha nodes are identified they are treated as the source from where information will be propagating through the network. We end up with X communities where X is the number of alpha nodes.

Random walk is a stochastic process describing a path that consists of random successive steps on some mathematical space. This method can also be used on graph data structure. The basic idea is that we have a walker that randomly explores a network, so a node having high visiting probability is considered to be near the central node of a community and can be considered to be part of that cluster. A method utilizing this approach, Multi-Walker Chain Bian et al. [12] is proposed where a group of K walkers is used. These walkers explore the network one by one for multiple iterations updating the visiting probability of each node. To identify the communities the top L nodes the largest mean scores is selected and the conductance value of the subgraph introduced by the nodes is calculated. Node set with the smallest value of conductance is returned as a community. Diffusion methods can be used to detect community structures in a network. In this type of approach a conceptual dye is injected on a particular node of multiple nodes in a network and watches the spread of the dye diffused over time steps across the edges of the network. The manner in which the dye diffuses provides us with hidden information on the structure of a graph. A Nonlinear Diffusion Ibrahim and Gleich [13] method for community detection is proposed where a semi-supervised technique is used, meaning that the diffusion adjusts by using feedback from the results. This method is repeated for a specified number of iterations or until there are no significant changes in the diffusion method.

C. Matrix Factorization

Matrix factorization (decomposition) is a collaborative filtering algorithm that works by decomposing the input matrix into two lower dimensional matrices making it easier to infer and calculate information from them. For example if we take number 10 then it can be factored into two parts, 2 and 5. Two most widely used methods are LU matrix decomposition and QR matrix decomposition.

Embedding is a process of representing a high dimensional non-linear structure like a graph into multiple 1d vectors each of same dimension preserving as much information possible from the original structure. The algorithm proposed in Skrlj et al. [14] traverses the embedding space and for each embedding tries multiple values of k (number of clusters). The method SCD uses two-step approach in finding the optimal value of k and further use it to identify communities in the network it does that based on a Silhouette score – SilhouetteGlobal (p, k) where p is the parameters passed to the embedding technique and k is the number of cluster to obtain in the network.

Hierarchical knowledge graphs can provide us with relevant real-world information about the clustering structure in a network, they are not always explicit in a network but can be useful in finding communities. We can decompose such HKGs to provide contextual information. This proposed approach Bhatt et al. [15] uses this knowledge graph to enhance the detection of communities by using the graph and the HKGs as input and finds community labels as well as the context from the HKG.

Normally graph embedding techniques and community detection are done separately. The proposed vGraph method Sun et al. [16] tries to solve both the problems by learning the embedding and detecting communities simultaneously by introducing a concept of node and community embedding also assuming that every node can be a part of multiple communities. Using this approach the representation of node can take advantage from the information gained by detection of node communities and vice versa.

Non negative matrix factorization (NFM) based methods factorize the adjacency matrix of a graph and convert it into two non-negative factor matrices. Now each column in the factor matrix can be analyzed as an inclination of a node to belong in different communities and the other can be used to identify mappings between the original network and the community membership. This proposed method Ye et al.

[17] Uses Deep Auto encoder like NMF for community detection which uses auto encoders to learn the mappings between the factor matrices. The components of the auto encoder guide each other in the learning phase obtaining an ideal community membership of nodes. This proposed method Adaptive Affinity Learning for Accurate Community Detection, Ye et al. [18] is another that uses Non negative matrix factorization (NFM). Using a technique to adaptively learn an affinity matrix and capture the essential equivalence between the nodes leading to improved community detection. This method embeds each node into a low-dimensional vector using transformation matrix, preserving the community structure. Using a mutual guiding system makes the model more accurate in detecting the similarity between the nodes.

D. Spectral Clustering

Spectral clustering is one of the oldest methods to detect clusters in graph dataset or real world non graph dataset. This clustering technique is inspired from graph theory. A graph can be represented in many different forms such as degree matrix, adjacency matrix or graph laplacian matrix. Spectral clustering clusters nodes on the basis of information gathered from eigenvalue of these matrices. The main contributions of spectral clustering in detecting communities are discussed below.

Fang Hu et al. [19] proposed a novel algorithm called node2vec-SC which consists of two phases. Node2vec is used to learn node embeddings of each node in the graph. Then spectral clustering technique is used to detect communities after calculating and finding eigenvalue and eigenvectors of similarity matrices, degree matrices and normalized laplacian matrices. This algorithm is also equally feasible for real world datasets.

Xiang Li et al. [20] discusses how spectral clustering can work on Heterogeneous Information Networks. Heterogeneous Information Networks (HINs) are the networks which are used to model real world interactions where every edge or link refers to a different type of interactions and relations between different types of vertices. Meta paths are a method of representing and modeling relations between different nodes in knowledge graphs or HINs. They propose a spectral clustering method where they form clusters by forming a similarity matrix with the help of Meta paths rather than random walks.

Spectral clustering being one of the oldest techniques to detect communities has one disadvantage that it is not scalable to large graph datasets and real world datasets because of its high computational complexity due to the formation of similarity matrix, graph laplacian and calculating eigenvalues and eigenvectors of this similarity matrix. Lingfei Wu et al. [21] use random binning features to speed up the process of formation of similarity matrix and calculation of eigenvalues and eigenvectors as they help in faster convergence. They also used single value decomposition (SVD) factorization method to calculate the eigenvalues faster.

Detecting overlapping communities is one the major challenges faced by researchers in the field of social network analysis. Yuan Zhang et al. [22] proposes an extended version of spectral clustering to solve the problem of overlapping com- munities. They have used K median technique for clustering eigenvalues of similarity matrix rather than using K means technique. This version of spectral clustering works well till communities are not largely overlapped.

Yixuan Li et al. [23] proposes a new spectral clustering method to detect communities in networks of large sizes. They propose two major changes in traditional spectral clustering. One of them is to initialize a random walk from seed nodes to detect nodes that may be present in target communities to reduce the number of computations of eigenvalues and eigenvectors. Other is to replace k-means algorithm in spectral space and find sparse vectors in to detect communities.

All these reviewed papers are mentioned in Table1 with all the datasets and techniques that were used. This table also lists all the different evaluation metrics that were being used by different researchers to assess their techniques such as NMI score, F1 score both of which are most popular evaluation metrics.

3. Discussion

Spectral clustering techniques are also used by various researchers to detect clusters or communities in graphs and networks. The advantage it has over other clustering techniques like k means is that spectral clustering does not presume any information about clusters or communities such k-means where it assumes that all the data points will be clustered around some centroids. The disadvantage it has shown is that it is very computationally expensive for large real world datasets because before clustering it has to make similarity matrix and calculate eigenvectors. Various researchers have used different techniques to speed up this process by using techniques such as SVD and random binning features.

Theoretically Neural networks can approximate any mathematical functions and possess great representation powers making it a very strong tool to work on graphs and networks. Neural networks architectures such as auto encoders, deep embedding techniques generates low level non linear embeddings which will be able to better represent the data in lower vector space and also able to preserve the non linear features of the data.

Matrix factorization techniques are able to decompose the matrix into smaller matrices which can improve the ability to detect the hidden features and ways the nodes are connected. One of the decomposed matrices can be used to identify one feature in the network and the others for a different feature and then use them in parallel to get better/accurate community detection results. These decomposed matrices are also able to preserve as much information as possible so we don’t have to worry about significant information loss as the result of decomposition.

Label propagation and random walk techniques have been proved to be very effective in detecting communities as they are able to go inside structure and propagate information at the node level. Label propagation is able to diffuse information from one node to the neighboring nodes and by analyzing the pattern of the diffusion of the labels we can gain important insight on the existing community structure in the network Similarly random walk algorithms walk on the network space from node to node learning information about the neighbors and thus aiding in the community detection process.

4. Conclusions

We live in a connected world on social media today, where forming groups is natural. These diversely sized and structured communities offer in-depth knowledge on how these systems interact and what attracts them to various groups. This framework clarifies the relationship between nodes and aids in our understanding of the environment, exposing social processes. In our study, we looked at the most recent papers on community detection and cutting-edge techniques. Based on community detection methods, we divided these publications into four subcategories. We examined well-known public datasets in addition to algorithms and methodologies. Research on social networks has grown recently and will continue to grow. There are various opportunities and challenges in community detection. It will be interesting to examine how community detection interacts with the graph representation learning and graph neural networks that have been the subject of much recent research. Other problems include making algorithms resistant to shifting graphs, detecting overlapping communities, and scaling on real-world networks. These challenges must be overcome and may open up new directions for investigation.

References

[1] Yang, Liang Cao, Xiaochun He, Dongxiao Wang, Chuan Wang, Xiao Zhang, Weixiong. (2016).

Modularity based community detection with deep learning.

[2] Dhilber, M. Surampudi, Durga. (2019). Community Detection in Social Networks Using Deep Learning. 10.1007/978-3-030-36987-3 15.

[3] Jia, Yuting Zhang, Qinqin Zhang, Weinan Wang, Xinbing. (2019). CommunityGAN: Community Detection with Generative Adversarial Nets.

[4] Kumar, Sanjay Panda, B Aggarwal, Deepanshu. (2020). Community detection in complex networks using network embedding and gravita- tional search algorithm. Journal of Intelligent Information Systems. 1-22. 10.1007/s10844- 020-00625-6.

[5] Cavallari, Sandro Zheng, Vincent Cai, Hongyun Chang, Kevin Cambria, Erik. (2017). Learning Community Embedding with Community Detection and Node Embedding on Graphs. 377-386. 10.1145/3132847.3132925.

[6] Chen, Z., Li, L., Bruna, J. (2019). Supervised Community Detection with Line Graph Neural Networks. arXiv: Machine Learning.

[7] Zhu, Xiaojin Ghahramani, Zoubin. (2003). Learning from Labeled and Unlabeled Data with Label Propagation.

[8] He, Kun Shi, Pan Bindel, David Hopcroft, John. (2019). Krylov Subspace Approximation for Local Community Detection in Large Networks. ACM Transactions on Knowledge Discovery from Data. 13. 52. 10.1145/3340708.

[9] Hollocou, Alexandre Bonald, Thomas Lelarge, Marc. (2018). Multiple Local Community Detection. ACM SIGMETRICS Performance Evalu- ation Review. 45. 76- 83. 10.1145/3199524.3199537.

[10] Martin Ester Hans-Peter Kriegel Jiirg Sander Xiaowei Xu.(1996) A Density-Based Algorithm for Discovering Clusters. Institute for Computer Science, University of Munich in Large Spatial Databases with Noise

[11] Venkatesaramani, Rajagopal Vorobeychik, Yevgeniy. (2018). Commu- nity Detection by Information Flow Simulation.

[12] Bian, Yuchen Ni, Jingchao Cheng, Wei Zhang, Xiang. (2017). Many Heads are Better than One: Local Community Detection by the Multi- walker Chain. 21-30. 10.1109/ICDM.2017.11.

[13] Ibrahim, Rania Gleich, David. (2019). Nonlinear Diffusion for Commu- nity Detection and Semi-Supervised Learning. WWW ’19: The World Wide Web Conference. 739-750. 10.1145/3308558.3313483.

[14] krlj, Blazˇ Kralj, Jan Lavrac, Nada. (2020). Embedding-based Silhouette community detection. Machine Learning. 109. 10.1007/s10994-020- 05882-8.

[15] Bhatt, Shreyansh Padhee, Swati Sheth, Amit Chen, Keke Shalin, Valerie Doran, Derek Minnery, Brandon. (2019). Knowledge Graph Enhanced Community Detection and Characterization. WSDM ’19: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. 51-59. 10.1145/3289600.3291031.

[16] Sun, Fan-Yun Qu, Meng Hoffmann, Jordan Huang, Chin-Wei Tang, Jian. (2019). vGraph: A Generative Model for Joint Community Detection and Node Representation Learning.

[17] Ye, Fanghua Chen, Chuan Zheng, Zibin. (2018). Deep Autoencoder- like Nonnegative Matrix Factorization for Community Detection. 1393- 1402. 10.1145/3269206.3271697.

[18] Ye, Fanghua Li, Shenghui Lin, Zhiwei Chen, Chuan Zheng, Zibin. (2018). Adaptive Affinity Learning for Accurate Community Detection. 1374-1379. 10.1109/ICDM.2018.00188.

[19] F. Hu, J. Liu, L. Li et al., Community detection in complex net- works using Node2vec with spectral clustering, Physica A (2019), doi: https://doi.org/10.1016/j.physa.2019.123633.

[20] Li, Xiang Kao, Ben Ren, Zhaochun Yin, Dawei. (2019). Spectral Clustering in Heterogeneous Information Networks.

[21] Wu, Lingfei Chen, Pin-Yu Yen, Ian Xu, Fangli Xia, Yinglong Aggarwal, Charu. (2018). Scalable Spectral Clustering Using Random Binning Features.

[22] Zhang, Y., Levina, E., Zhu, J. (2020). Detecting Overlapping Commu- nities in Networks Using Spectral Methods. SIAM J. Math. Data Sci., 2, 265-283.

[23] Li, Y. He, Kun Bindel, David Hopcroft, J.E.. (2015). Uncovering the small community structure in large networks: a local spectral approach. Proceedings of the 24th international conference on world wide web. 658-668.

[24] https://en.wikipedia.org/wiki/File:Network Community Structure.svg

[25] https://www.researchgate.net/figure/Zacharys-karate-club- network-Members-of-the-communities-resulting- after-the-split- are fig2 225168779

[26] https://www.researchgate.net/figure/Sample-LFR-benchmark-graph-of- size-n-100-with-parameter-values- set-to-M-axDeg-30 fig8 325719324

[27] https://towardsdatascience.com/a-tale-of-two-convolutions-differing- design-paradigms-for-graph- neural-networks-8dadffa5b4b0