Statistical and Machine Learning Methods for Networks

by

Huimin Cheng

(Under the Direction of Wenxuan Zhong and Ping Ma)

Abstract

In this thesis, we focus on developing novel statistical and machine learning methods for network data analysis, with emphasis on both appealing statistical properties and computational efficiency. In particular, the first project studies the problem of network cross-validation in graphon estimation. Graphon, short for graph function, provides a generative model for networks. The success of most graphon estimation methods depends on a proper specification of hyperparameters. Existing network cross-validation methods suffer from restrictive model assumptions, expensive computational costs, and a lack of theoretical guarantees. To address these issues, we propose a masked mirror validation (MMV) method. The second project studies the problem of network sampling. In the past decades, many large graphs with millions of nodes have been collected/constructed. The high computational cost and significant visualization difficulty hinder the analysis of large graphs. To overcome the computational challenge of a large graph, we propose a graph subsampling algorithm, i.e., Ollivier-Ricci curvature Gradient-based subsampling (ORG-sub) algorithm, which employs Riemannian geometric information. The superiority of the proposed methods is demonstrated by various synthetic and real experiments. The third project developed and applied network analysis methods to analyze transnational advocacy networks (TANs). We build a dataset of the 3,903 NGOs connected through 1.3 million ties occurring through meetings and conferences for NGOs put on or coordinated by the United Nations. Using community detection methods, we identify

four distinct communities in the overall NGO network, with differences in distributions of brokerage roles across communities. This help us better understand how the TANs simultaneously provides social power and exacerbates global inequalities.

INDEX WORDS: Network Cross-Validation; Network Sampling; Network Applications

Statistical and Machine Learning Methods for Networks

by

Huimin Cheng

B.S., Hubei University of Economics, Wuhan, China, Class of 2015

M.S., Central University of Finance and Economics, Beijing, China, Class of 2017

A Dissertation Submitted to the Graduate Faculty of the

University of Georgia in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2023

Statistical and Machine Learning Methods for Networks

by

Huimin Cheng

Major Professor:

Wenxuan Zhong

Co-major Professor:

Ping Ma

Committee:

Amanda Murdie

Bingqian Xu

Electronic Version Approved:

Ron Walcott

Dean of the Graduate School

The University of Georgia

May 2023

Acknowledgments

This work was made possible through the support of many people. I cannot thank my advisors Prof. Wenxuan Zhong and Prof. Ping Ma enough, who have generously provided me with assistance, guidance, and support throughout my Ph.D. study. I have been greatly impressed and influenced by their insights, wisdom, and passion for statistics. I would like to express my deepest gratitude to them. Meanwhile, I also want to express my sincere gratitude to my committee members: Prof. Bingqian Xu and Prof. Amanda Murdie, for their advice, help, and encouragement on my dissertation research. It is my honor to have them all on my committee.

I want to express my thanks to all labmates of Big Data Analytics Lab (BDAL) for their support: Xiaoxiao Sun, Yiwen Liu, Xin Xing, Xinlian Zhang, Wei Xu, Rui Xie, Cheng Meng, Jingyi Zhang, Ye Wang, Huimin Cheng, Jun Yu, Yongkai Chen, Shushan Wu, Zhen Wang, Haoran Lu, Jiazhang Cai, Luyang Fang, Tao Wang, Yufang Liu and many others whose names are not listed. It was fantastic to have the opportunity to work with these young researchers in the past five years. I would also like to express my genuine gratitude to department head Prof. T.N. Sriram, graduate coordinator Prof. Liang Liu, and all members of the Department of Statistics and at the University of Georgia. They have truly influenced me through their extraordinary research experiences. Additionally, I benefited tremendously from intellectual discussions with colleagues from Prof. Jun Yu, Prof. Shen Ye, Prof. Wenzhan Song, and Prof. Guocheng Yuan.

Part of the dissertation research is supported by U.S. National Science Foundation grants DMS-1903226, DMS-1925066, DMS-2124493 and NIH grants R01GM1222080a to Prof. Ma and Prof. Zhong.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the past two decades, networks (a.k.a. graphs) have shown surprising effectiveness in characterizing complex systems (Barabási, 2016; Kolaczyk et al., 2020). A network consists of a collection of nodes and edges, where a node in a network is a system component, and an edge is an interaction between components (Rohe et al., 2011). Many real-world data come naturally in the form of a network (a.k.a. graph), e.g., social network modeling the friendship relationship between individuals, gene expression network modeling the co-expression relationship between genes. Fueled by the rapid development of the internet and telecommunication technology, network (a.k.a. graph) analysis is becoming a major technology for revealing relevance, extracting knowledge, and making decisions. It not only transforms our daily behaviors, including working, collaborating, entertaining, health monitoring and etc, but also leads to scientific breakthroughs and even maintains people's or country's relationships (Fan et al., 2019; M. Newman, 2018).

Despite the progress in network data analysis, there are still many challenges and open questions, such as inference, prediction, and scalability, which require further research and the development of novel methods and tools. My doctoral research goal is to develop novel theoretically justifiable and computationally efficient methods for the network data. In particular, we investigated the following problems in network analysis.

1. In Chapter 2, we developed a network cross-validation method, i.e., masked mirror validation, to select the optimal hyperparameter when estimating graphon model parameters.
2. In Chapter 3, we developed a network sampling method, i.e., Ollivier-Ricci curvature Gradient-based subsampling (ORG-sub), to answer a key question: "Given a huge real network, can we derive a representative subsample that preserves key structures?".
3. In Chapter 4, we analyzed transnational advocacy networks that represents international relations, to investigate how the transnational advocacy network simultaneously provides social power and exacerbates global inequalities.

1.1 Masked Mirror Validation in Graphon Estimation

Statistical analysis of networks has gained significant attention due to its versatile applications in fields such as sociology, physics, biology, and medical sciences. By employing statistical models, researchers can efficiently extract crucial structural information from the network while filtering out noisy and uninformative details. Mathematically, the connectivity of a network (a set of nodes and edges) with n nodes can be modeled by an $n \times n$ adjacency matrix $\mathbf{A}$ with ijth entry a_{ij}, where $a_{ij} = 1$ if node i is connected with node j and $a_{ij} = 0$ otherwise. To understand the generative mechanism of a network, researchers usually assume that a_{ij}s are independently distributed from a Bernoulli distribution with mean p_{ij}, i.e.,

$$a_{ij} \overset{\text{ind}}{\sim} \text{Ber}(p_{ij}). \tag{1.1}$$

Clearly, p_{ij} is not estimable as there is only one observation for each parameter p_{ij}. To make p_{ij} estimable, researchers further assume

$$p_{ij} = f(\mu_i, \mu_j), \tag{1.2}$$

where $f : [0, 1] \times [0, 1] \to [0, 1]$ is referred to as the graphon function, μ_i and μ_j are unknown latent positions of node i and j. Eq. (2.1) and (2.2) together are referred to as graphon model (Lovász & Szegedy, 2006). Without loss of generality, we further assume that f is a bounded symmetric function and moreover

μ_i and μ_j are independently and identically distributed from Uniform [0,1] distribution. The graphon model is arguably the most general graph model of which many popular graph models can be considered as its special cases, including Erdős–Rényi random graph model (Gilbert, 1959) that assumes a constant function and the stochastic block model (SBM) (Holland et al., 1983) that assumes a step function.

An accurate estimation of graphon plays a key role in many applications, such as link prediction. In recent decades, various methods for graphon estimation have been proposed. The success of most graphon estimation methods depends on a proper specification of hyperparameters. Conventional cross-validation methods for hyperparameter tuning cannot be directly applied to network data. Some network cross-validation methods have been proposed, but they suffer from restrictive model assumptions, expensive computational costs, and a lack of theoretical guarantees.

To address these issues, we propose a masked mirror validation (MMV) method in Chapter 2. A distinctive feature of MMV is that instead of splitting data into training and validation, MMV constructs the training network which is analogous to a mirror "image" of the original network, but with a subgraph being masked by a specifically randomly generated subgraph. The masked subgraph is used for validation. Specifically, MMV first constructs two independent networks for training and validation. The training network is analogous to a mirror "image" of the original network, but with a subgraph being masked by a specifically randomly generated subgraph. The masked subgraph is used for validation. MMV then uses the validation data to evaluate the accuracy of the graphon estimator. Asymptotic properties of the MMV are established. The effectiveness of the proposed method in terms of both computation and accuracy is demonstrated by extensive simulation studies and real experiments. In sum, compared with existing methods, MMV enjoys (1) rigorous theoretical underpinning, (2) efficient computation, and (3) superior estimation accuracy.

1.2 Network Sampling Using Ricci Curvature

In the past decades, many large graphs with millions of nodes have been collected/constructed. For example, (Wang et al., 2011) presented a Twitter social network, which has more than 190 million nodes

(users) who generate more than 65 million edges (tweets) every day (Wang et al., 2011). Such huge networks enable researchers to tackle more complex problems. However, they pose great challenges to storing, visualizing, and analyzing since their sheer volumes render many computational methods infeasible. To overcome the computation difficulties, researchers have developed many graph subsampling approaches to provide a rough sketch that preserves global properties. Graph subsampling has emerged as a powerful technique for efficiently analyzing large-scale network data. By selecting a smaller subset of nodes and edges from the original graph, subsampling can reduce the computational cost and memory requirements of network analysis while preserving the key structural properties and patterns of the network. These graph subsampling methods can help researchers estimate the graph statistics, e.g., the number of communities, of the large graph from the subsample.

Various graph subsampling methods that preserve different graph properties have been proposed, including node sampling (Mall et al., 2013; Zeng et al., 2019), edge sampling (Krishnamurthy et al., 2005), and exploration sampling (Goodman, 1961; Hübler et al., 2008; Leskovec et al., 2005; Maiya & Berger-Wolf, 2010). Random sampling is the simplest and most widely used subsampling method, where nodes and edges are selected uniformly at random from the original graph. Edge-based sampling involves selecting edges based on their importance or weight, such as selecting the top k edges with the highest betweenness centrality or edge weight. Node-based sampling involves selecting nodes based on their importance or relevance to the network, such as selecting the top k nodes with the highest degree or eigenvector centrality. These methods can be used in combination to achieve different goals and trade-offs, such as selecting nodes with high degrees and edges with high betweenness centrality to preserve the network's community structure.

An important graph feature that has received less attention is the number of communities (denoted by M), which plays a crucial role in identifying the community structures. Network data often have natural communities, and the identification of these communities helps answer vital questions in a variety of fields (Rohe et al., 2011). For example, communities in social networks may represent groups of people who share a similar interest, and communities in protein-protein interaction networks could be regulatory

modules of interacting proteins (Rohe et al., 2011). Despite many applications of existing subsampling methods, they tend to leave out minority communities because nodes with high degrees are more likely to be sampled.

To overcome the shortcomings of the existing methods, we are motivated to apply the community information hidden in the graph to the subsampling method. Though the community structure is unavailable, community structure information can be obtained by applying geometric methods to a graph. An analog of Ricci curvature in the manifold is defined for the graph, i.e., Ollivier Ricci curvature. Based on the asymptotic results about the within-community edge and between-communities edge's OR curvature, we propose a subsampling algorithm based on our theoretical results, the Ollivier-Ricci curvature Gradient-based subsampling (ORG-sub) algorithm in Chapter 3. The proposed ORG-sub algorithm has two main contributions: First, ORG-sub provides a rigorous theoretical guarantee that the probability of ORG-sub taking all communities into the final subgraph converges to one. Second, extensive experiments on synthetic and benchmark datasets demonstrate the advantages of our algorithm.

1.3 Analysis of Transnational Advocacy Network in Political Science

Transnational advocacy networks (TANs) are the most common example of networks in international relations. Despite their familiarity, we know little about how advocacy networks of nongovernmental organizations (NGOs) are structured. Drawing on the cross-disciplinary concepts of emergent communities and distinct brokerage roles, we argue that the network may reinforce power disparities and inequalities at the very same time that it provides social power. TANs are similar to emergent communities of practice, with some organizations acting as various types of brokers within and between communities. Preexisting resources are more likely to lead global North organizations to occupy brokerage roles that provide additional agenda-setting and resource-allocating power. We build a dataset of the 3,903 NGOs connected through 1.3 million ties occurring through meetings and conferences for NGOs put on or coordinated

by the United Nations. Using community detection methods, we identify four distinct communities in the overall NGO network, with differences in distributions of brokerage roles across communities. Examining the communities, brokerage role distributions, and preexisting power disparities can help us better understand the divergent findings in previous literature and conceptualize TANs

Chapter 2

Masked Mirror Validation in Graphon Estimation

2.1 Introduction

Fueled by the rapid development of the internet and telecommunication technology, network (a.k.a. graph) analysis is becoming a major technology for revealing relevance, extracting knowledge, and making decisions. The connectivity of a network (a set of nodes and edges) with n nodes can be modeled by an $n \times n$ adjacency matrix $\mathbf{A}$ with ijth entry a_{ij}, where $a_{ij} = 1$ if node i is connected with node j and $a_{ij} = 0$ otherwise. In this paper, we focus on a single network setting, i.e., we only observe one network. To understand the generative mechanism of a network, researchers usually assume that a_{ij}s are independently distributed from a Bernoulli distribution with mean p_{ij}, i.e.,

$$a_{ij} \overset{\text{ind}}{\sim} \text{Ber}(p_{ij}). \tag{2.1}$$

Clearly, p_{ij} is not estimable as there is only one observation for each parameter p_{ij}. To make p_{ij} estimable, researchers further assume

$$p_{ij} = f(\mu_i, \mu_j), \tag{2.2}$$

where $f : [0,1] \times [0,1] \to [0,1]$ is referred to as the graphon function, μ_i and μ_j are unknown latent positions of node i and j. Eq. (2.1) and (2.2) together are referred to as graphon model (Lovász & Szegedy, 2006).

Graphon model is arguably the most general network model. Many popular network models can be cast as its special cases, including Erdős–Rényi random graph model (Gilbert, 1959) where f is a constant, and the stochastic block model (SBM) (Holland et al., 1983) where f is a step function. Unfortunately, f is not identifiable as f and μ_is are confounding with each other (Diaconis & Janson, 2007). To address this identifiability issue, one line of research imposes strong assumptions on f. For example, Chan and Airoldi, 2014 assumes that the expected degree $\int_0^1 f(\mu_i, \mu_k) d\mu_k$ is strictly monotone. In this way, μ_i can be represented by the corresponding node degree, thus solving the identifiability problem. Nevertheless, this assumption is strong, e.g., it excludes the stochastic block model.

Another line of research to address this identifiability issue focuses on estimating the edge connecting probability matrix $\mathbf{P}_0 = (p_{ij})$ instead of f. As pointed out by Zhang et al., 2017, in practice, the main purpose of estimating f is to estimate $\mathbf{P}_0$, thus lack of identifiability of f may not matter if $\mathbf{P}_0$ itself can be estimated. Even without knowing the μ_is, it is still possible to estimate the matrix $\mathbf{P}_0$ under the graphon model structure. The estimation of $\mathbf{P}_0$ is also called graphon estimation in literature (Gao et al., 2015). Various methods have been proposed for estimating $\mathbf{P}_0$ in the past decade. For example, a low-rank approximation method was proposed to estimate $\mathbf{P}_0$ by imposing a low-rank assumption on $\mathbf{P}_0$ (Chatterjee, 2015); a neighborhood smoothing method was proposed by assuming that f comes from a piece-wise Lipschitz family (Qin et al., 2021; Zhang et al., 2017). Comprehensive reviews of this line of research can be found in Chandna et al., 2021; Gao et al., 2015.

Despite the overall appealing performance, the effectiveness of graphon estimation methods highly depends on the selection of hyperparameters. How to select the optimal hyperparameter that leads to the most accurate graphon estimation remains elusive. Conventional cross-validation (CV) methods (Arlot & Celisse, 2010; Stone, 1974; Sun et al., 2021; Wahba, 1990) can not be directly applied due to two facts. First, we only have a single network observation. Second, even though we can treat each node pair in the

network as a replication, holding out a set of node pairs as validation data will make the leftover training network has missing values (Li et al., 2020).

We are thus in urgent need of a validation method for hyperparameter tuning in graphon estimation. To the best of our knowledge, research along this line is still lacking. Some early work includes: E. M. Airoldi et al., 2008 employed node-pair splitting method under a Bayesian network framework; Chen and Lei, 2018 developed a network cross-validation criterion by node sampling for determining the optimal number of communities. Nevertheless, these methods are restricted to SBM and a specific task in SBM (i.e., determining the number of communities). More recently, Li et al., 2020 proposed an edge cross-validation (ECV) method which does not assume a specific form of the graphon function. Despite many successful applications, ECV suffers from three practical challenges. (1) Restrictive assumption: ECV requires $\mathbf{P}_0$ to be low-rank. (2) Intensive computation: ECV implements the matrix completion for an $n \times n$ adjacency matrix in each cross-validation fold, which is expensive. (3) A lack of theoretical guarantee: the theoretical properties of ECV in hyperparameter selection for graphon estimation remain elusive.

Masked Mirror validation. To address the aforementioned issues, we propose a $K-$fold masked mirror validation (MMV) method. In each fold, MMV first constructs the training data $\tilde{\mathbf{A}}$ by constructing a mirror "image" of $\mathbf{A}$ with a subset being randomly masked. In $\tilde{\mathbf{A}}$, entries without masks are the same as those of $\mathbf{A}$. In $\tilde{\mathbf{A}}$, entries with masks are randomly independently generated from a specific Bernoulli distribution. The aforementioned masked subset is used for validation. In this way, $\tilde{\mathbf{A}}$ only has the information of the unmasked subset in $\mathbf{A}$, and the validation data only has the information of the masked subset in $\mathbf{A}$. Thus, training data and validation data are independent, which is critical to CV. MMV then uses the training data $\tilde{\mathbf{A}}$ to get an estimator, which is shown to be a biased estimator of $\mathbf{P}_0$. We hence correct the bias to get an unbiased estimator. We then evaluate the performance of the graphon estimation on the validation data by calculating a MMV score. Specifically, MMV score is calculated by comparing the masked entries in $\mathbf{A}$ and the entries in the debiased estimator. Finally, MMV outputs the hyperparameter that minimizes the averaged MMV score over K folds as the best candidate.

Our proposed MMV enjoys three advantages over existing methods. First, MMV does not impose any assumption on the graphon function form and is thus broadly applicable. Second, MMV has rigorous theoretical underpinning. Particularly, we use the MSE between the estimator and the true probability matrix as a gold standard metric to evaluate the estimation error. We theoretically prove that the minimizer of the MMV asymptotically converges to the minimizer of the MSE. Third, MMV is fast and can be easily applied to large networks. Compared with ECV, MMV does not perform expensive matrix completion, thus significantly accelerating the procedure. Forth, our extensive simulation studies and real data applications, show that MMV is a promising framework for hyperparameter selection.

The rest of the paper is organized as follows. In Section 2.2, we give a brief introduction to the model setup, followed by a presentation of our proposed MMV method. The asymptotic behavior of the MMV is discussed in Section 2.3. Simulation and real data examples are reported in Section 2.5 and Section 2.6. Additional remarks in Section 2.7 conclude the paper.

2.2 Masked Mirror Validation

Recall that $\mathbf{A} = (a_{ij})$ denotes the adjacency matrix of a network generated by graphon model Eq.(2.1) and (2.2) with parameter $\mathbf{P}_0 = (p_{ij})$. If $a_{ij} = a_{ji}$ for any i and j, a network is called an undirected network. Otherwise, this network is called a directed network. Our proposed method is applicable to both undirected and directed networks. For ease of presentation, we only present the algorithm for the undirected network in this paper. The algorithm for the directed network can be analogously derived by treating node pairs v_iv_j and v_jv_i as two different pairs.

2.2.1 Mean Squared Error

Let M denote the hyperparameter, $\mathbf{P}_M^{\mathbf{A}} = (\hat{p}_{ij})$ denote the estimator of $\mathbf{P}_0$ using M and data $\mathbf{A}$. To measure the estimation error of $\mathbf{P}_M^{\mathbf{A}}$, we consider the mean squared error (MSE), which is defined as

$$\text{MSE}(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0) = \frac{1}{n(n-1)/2} \sum_{i<j} (\hat{p}_{ij} - p_{ij})^2. \tag{2.3}$$

Let $\mathcal{M} = \{M_1, \ldots, M_m\}$ denote the candidate pool consisting of m candidate hyperparameters. Note that similar to many existing CV methods (Chen & Lei, 2018; Li et al., 2020), $\mathcal{M}$ is predefined by users, and is a finite set of discrete values. Given $\mathcal{M}$, we aim to find the optimal hyperparameter M_o^{mse} by minimizing the MSE, i.e.,

$$M_o^{mse} = \text{argmin}_{M \in \mathcal{M}} \text{MSE}(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0). \tag{2.4}$$

Minimizing MSE is equivalent to minimizing another population version objective function which motivates a CV procedure, as we will show below. We define

$$L(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0) = \frac{1}{n(n-1)/2} \sum_{i<j} \mathbb{E}_{\mathcal{A}_{ij}}[(\hat{p}_{ij} - \mathcal{A}_{ij})^2 | \mathbf{A}],$$

where $\mathcal{A}_{ij}$s are independent random variables following $\mathcal{A}_{ij} \sim \text{Ber}(p_{ij})$. In particular, $L(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0)$ is a conditional expectation given $\mathbf{A}$. Note that given $\mathbf{A}$, $\hat{p}_{ij}$ is also given, since $\hat{p}_{ij}$ is obtained from $\mathbf{A}$. We

rewrite $L(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0)$ as

$$\begin{aligned} L(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0) &= \frac{1}{n(n-1)/2} \sum_{i<j} \mathbb{E}_{\mathcal{A}_{ij}}[(\hat{p}_{ij} - p_{ij} + p_{ij} - \mathcal{A}_{ij})^2 | \mathbf{A}] \\ &= \frac{1}{n(n-1)/2} \sum_{i<j} \left\{ \mathbb{E}_{\mathcal{A}_{ij}}[(\hat{p}_{ij} - p_{ij})^2 | \mathbf{A}] + \mathbb{E}_{\mathcal{A}_{ij}}[(p_{ij} - \mathcal{A}_{ij})^2 | \mathbf{A}] \right\}. \\ &= \frac{1}{n(n-1)/2} \sum_{i<j} \left\{ (\hat{p}_{ij} - p_{ij})^2 + p_{ij}(1 - p_{ij}) \right\} \\ &= \text{MSE}(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0) + \frac{1}{n(n-1)/2} \sum_{i<j} p_{ij}(1 - p_{ij}). \end{aligned}$$

Since the second term does not depend on M, minimizing MSE is equivalent to minimizing $L(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0)$. Thus, by finding the minimizer of $L(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0)$, we can find M_o^{mse}.

Next, we show the intuition of applying CV to estimate $L(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0)$. Let us imagine an ideal scenario where we have other K networks, whose adjacency matrices are $\mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(K)}$ which are independently and identically generated by the graphon model in Eq. (2.1) with parameter $\mathbf{P}_0 = (p_{ij})$. Note that in reality, we only have one observation $\mathbf{A}$ and do not have $\mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(K)}$. We only use this hypothetical scenario to illustrate intuition. Under this hypothetical case, the sample-based estimation of $L(\mathbf{P}_M, \mathbf{P}_0)$ is $\frac{1}{Kn(n-1)/2} \sum_{k=1}^{K} \sum_{i<j} (\hat{p}_{ij} - b_{ij}^{(k)})^2$, where $b_{ij}^{(k)}$ is the ijth element of $\mathbf{B}^{(k)}$. Calculating this estimator is equivalent to the following cross-validation procedure: We use $\mathbf{A}$ as the training data to obtain $\mathbf{P}_M^{\mathbf{A}}$; We then use those K hypothetical networks as validation networks; We finally calculate the loss in validation data. This motivates us to use the CV procedure to estimate $L(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0)$ and M_o^{mse}.

Unfortunately, we only have a single observation $\mathbf{A}$. To address this issue, a natural solution is to split the network $\mathbf{A}$ into two subgraphs: One subgraph is used for training, and the rest is used for validation. Nevertheless, as pointed out by Li et al., 2020, when sampling a subgraph as validation data, the leftover network has missing values, which means many models and methods cannot be applied to it directly.

2.2.2 Masked Mirror Validation Procedure

To address the aforementioned challenge, we propose the masked mirror validation procedure that selects the optimal hyperparameter with a theoretical guarantee. Let $\mathcal{S}_{\mathbf{A}}$ denote the set of all node pairs in $\mathbf{A}$, the number of elements in $\mathcal{S}_{\mathbf{A}}$ is $|\mathcal{S}_{\mathbf{A}}| = \frac{n(n-1)}{2}$.

Our proposed $K-$fold MMV works in the following steps. MMV first randomly splits $\mathcal{S}_{\mathbf{A}}$ into K equal-sized subsets $\mathcal{S}_1, \ldots, \mathcal{S}_K$. For $k = 1, \ldots, K$, and for each $M \in \mathcal{M}$, MMV then repeats the following steps to get the loss on the kth validation data.

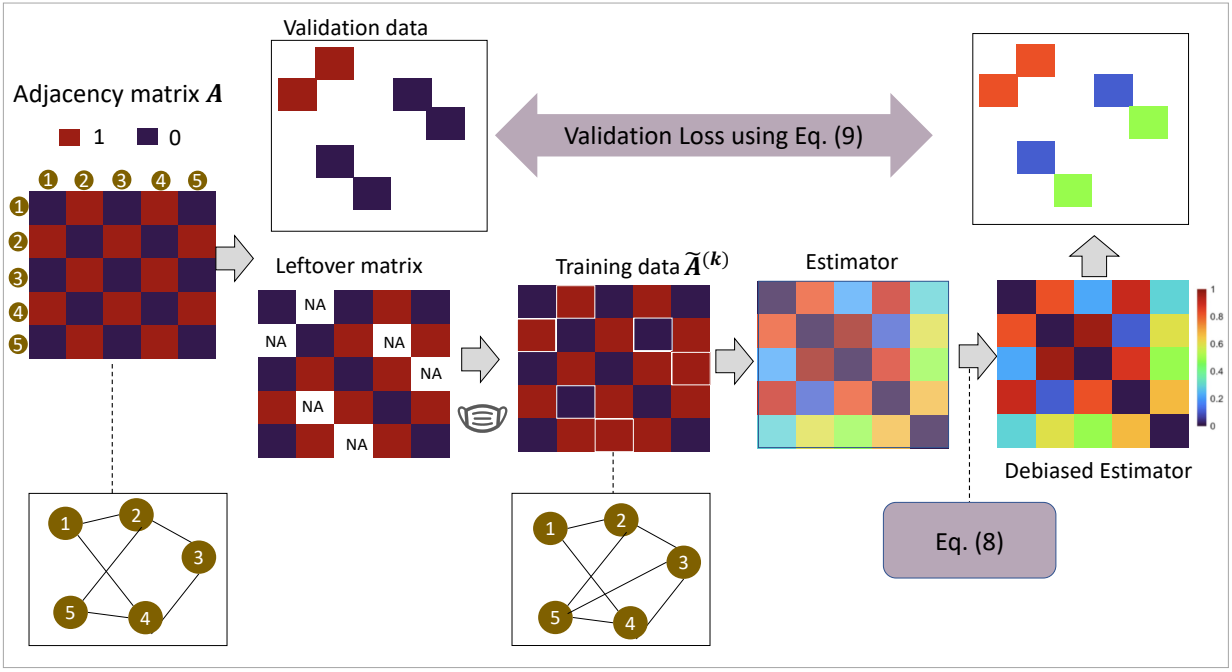


Figure 2.1: Illustration of steps 1-4 of MMV using a toy example of an undirected network with five nodes. The left heatmap shows the adjacency matrix of a five-node network. Three node pairs (v_1v_2, v_2v_4, v_3v_5) are randomly selected for validation. The leftover matrix has holes. We mask these holes and generate the training matrix $\tilde{\mathbf{A}}^{(k)}$. In this example, the difference between $\tilde{\mathbf{A}}^{(k)}$ and $\mathbf{A}$ is that node pair v_3v_5 has no edge in $\mathbf{A}$, while v_3v_5 has an edge in $\tilde{\mathbf{A}}^{(k)}$. Then we get $\mathbf{P}_M^{\tilde{\mathbf{A}}^{(k)}}$ using the training matrix. By correcting the bias using Eq. (2.8), we get an adjusted estimator. The validation loss is calculated by comparing the masked entries in $\mathbf{A}$ and the entries in the adjusted estimator.

Step 1. Select the kth subset $\mathcal{S}_k$ as the validation data. The selected subset is removed from $\mathbf{A}$. A toy example is shown in Figure 2.1. The left heatmap shows the adjacency matrix $\mathbf{A}$ of an undirected network

with five nodes. Three node pairs, i.e., v_1v_2, v_2v_4 and v_3v_5, are selected into the validation set. These three node pairs are thus missing in the leftover matrix, making the leftover data look like a matrix with holes.

Step 2. We mask these holes by randomly generating binary entries using $\text{Ber}(\theta)$, where $\theta \in [0, 1]$ is a predefined parameter. The new matrix with masks is our training data $\tilde{\mathbf{A}}^{(k)} = (\tilde{a}_{ij}^{(k)})$. Mathematically,

$$\tilde{a}_{ij}^{(k)} = \begin{cases} a_{ij}, & \text{if } \mathrm{v}_i v_j \in \mathcal{S}_{-k}, \\ \tilde{b}_{ij}, \text{ where } \tilde{b}_{ij} \overset{i.i.d}{\sim} \text{Ber}(\theta) & \text{if } \mathrm{v}_i v_j \in \mathcal{S}_k, \end{cases} \tag{2.5}$$

where a_{ij} is the ijth entry of $\mathbf{A}$. In particular, when $\theta = 0$, we mask the holes of the leftover matrix with all zeros; when $\theta = 1$, we mask the holes with all ones. We relegate the discussion of how to set θ to Section 2.4. In the toy example of Figure 2.1, MMV masks v_1v_2 with one, masks v_2v_4 with zero, and masks v_3v_5 with one. The training data is analogous to a mirror "image" of $\mathbf{A}$ with a subset of entries masked by noises.

Since $\tilde{a}_{ij}^{(k)}$ is obtained by adding noises using Eq. (2.5), it is expected that the distributions of a_{ij} and $\tilde{a}_{ij}^{(k)}$ are not identical. In particular, we employ the law of total probability to derive the probability of $\tilde{a}_{ij}^{(k)} = 1$, that is,

$$\begin{aligned} \mathbb{P}(\tilde{a}_{ij}^{(k)} = 1) &= \mathbb{P}(\tilde{b}_{ij} = 1 | v_i v_j \in \mathcal{S}_k)\mathbb{P}(v_i v_j \in \mathcal{S}_k) + \mathbb{P}(a_{ij} = 1 | v_i v_j \in \mathcal{S}_{-k})\mathbb{P}(v_i v_j \in \mathcal{S}_{-k}) \quad (2.6) \\ &= \frac{\theta}{K} + \frac{(K-1)p_{ij}}{K}, \quad (2.7) \end{aligned}$$

where K and θ are both known. Fortunately, there is a one-to-one map between $\mathbb{P}(\tilde{a}_{ij}^{(k)} = 1)$ and p_{ij}. Thus we can infer p_{ij} using estimated $\mathbb{P}(\tilde{a}_{ij}^{(k)} = 1)$.

Step 3. We apply M to training data $\tilde{\mathbf{A}}^{(k)}$ to get $\mathbf{P}_M^{\tilde{\mathbf{A}}^{(k)}}$, of which the ijth entry is an estimator of $\mathbb{P}(\tilde{a}_{ij}^{(k)} = 1)$. As the previous analysis shows, there is a one-to-one map between $\mathbb{P}(\tilde{a}_{ij}^{(k)} = 1)$ and p_{ij}. We then use an inverse map to derive an estimator of p_{ij}. From Eq. (2.7), we have $p_{ij} = \frac{K\mathbb{P}(\tilde{a}_{ij}^{(k)}=1)-\theta}{K-1}$. Thus, we get the new estimator $\overset{*}{\mathbf{P}}_M^{\tilde{\mathbf{A}}^{(k)}} = (\overset{*}{p}_{ij}^{(k)})$ where

$$\overset{*}{\mathbf{P}}_M^{\tilde{\mathbf{A}}^{(k)}} = \frac{K\mathbf{P}_M^{\tilde{\mathbf{A}}^{(k)}} - \theta}{K-1}. \tag{2.8}$$

Step 4. We calculate the validation loss, i.e., by comparing the validation data with the masked part in $\overset{*}{\mathbf{P}}_M^{\tilde{\mathbf{A}}^{(k)}}$. The kth validation loss is defined as

$$L_M^{(k)}(\mathbf{A}) = \frac{1}{|\mathcal{S}_k|} \sum_{v_i v_j \in \mathcal{S}_k} (\overset{*}{p}_{ij}^{(k)} - a_{ij})^2, \tag{2.9}$$

where $|\mathcal{S}_k|$ is the number of node pairs in $\mathcal{S}_k$.

By repeating steps 1-4 for $k = 1, \ldots, K$, we calculate the MMV score of M by averaging all validation losses,

$$\text{MMV}_M(\mathbf{A}) = \frac{1}{K} \sum_{k=1}^{K} L_M^{(k)}(\mathbf{A}). \tag{2.10}$$

A smaller score indicates a smaller estimation error using M. Finally, we select the candidate with the smallest MMV score, i.e., choose

$$M_o = \text{argmin}_{M \in \mathcal{M}} \text{MMV}_M(\mathbf{A}). \tag{2.11}$$

We summarize our MMV method in Algorithm 1.

2.2.3 Computational Complexity

The complexity of the Algorithm 1 can be analyzed by considering each step individually. For each $i = 1, \ldots, m$ and $k = 1, \ldots, K$, constructing the training network $\tilde{\mathbf{A}}^{(k)}$ requires $O(|\mathcal{S}_k|)$ computation. The computation of obtaining $\mathbf{P}_{M_i}^{\tilde{\mathbf{A}}^{(k)}}$ depends on the specific graphon estimation method, whose computational cost is denoted by $O(r)$. Getting the debiased estimator and calculating the validation loss requires $O(n^2)$ computation. Thus, the computational cost of MMV is $O(Km \times \max(r, n^2)))$, where K is the number of folds, and m is the number of candidate hyperparameters. In comparison, the computational complexity of ECV is $O(Km \times \max(r, n^3))$. When applying MMV to tune hyperparameters

Algorithm 1 MMV for parameter selection or method selection

Input: (1) The adjacency matrix of a single observed graph $\mathbf{A}$; (2) A set $\mathcal{M}$ of m candidate methods or hyperparameters; (3) Two tuning parameters θ and K.
Splitting step. Randomly split all node pairs $\mathcal{S}_{\mathbf{A}}$ into K equal-sized sets, $\mathcal{S}_1, \ldots, \mathcal{S}_K$.
for $i = 1, \ldots, m$ **do**
  for $k = 1, \ldots, K$ **do**
    (a) Construct training network $\tilde{\mathbf{A}}^{(k)}$ by Eq. (2.5).
    (b) Apply M_i to $\tilde{\mathbf{A}}^{(k)}$ to get $\mathbf{P}_{M_i}^{\tilde{\mathbf{A}}^{(k)}}$.
    (c) Get debiased estimator $\overset{*}{\mathbf{P}}_{M_i}^{\tilde{\mathbf{A}}^{(k)}}$ using Eq. (2.8)
    (d) Apply Eq. (2.9) to get the loss of validation data $L_{M_i}^{(k)}(\mathbf{A})$ in kth replication.
  end for
  Calculate MMV score according to Eq. (2.10)
end for
Output: The hyperparameter with the smallest MMV score.

for a method whose computational complexity is less than $O(n^2)$, such as SBA, the computational cost of MMV is $O(Kmn^2)$, while that of ECV is $O(Kmn^3)$. In this case, MMV is much faster than ECV.

2.3 Theoretical Properties

To build the selection consistency property of MMV, i.e., M_o in Eq. (2.11) is asymptotically consistent with M_o^{mse} in Eq. (2.4), we need to impose the following assumptions.

Assumption 1. $K \to \infty$ as $n \to \infty$.

Assumption 2. For any $M \in \mathcal{M}$ and any $k \in \{1, \ldots, K\}$, we have $\|\mathbf{P}_M^{\tilde{\mathbf{A}}^{(k)}} - \mathbf{P}_M^{\mathbf{A}}\|_F / \|\tilde{\mathbf{A}}^{(k)} - \mathbf{A}\|_F = O_p(1)$.

Assumption 1 requires that K increases when n increases. In particular, leave-one-out is an example that satisfies Assumption 1. Assumption 2 essentially requires that the ratio between $\|\mathbf{P}_M^{\tilde{\mathbf{A}}^{(k)}} - \mathbf{P}_M^{\mathbf{A}}\|_F$ and $\|\tilde{\mathbf{A}}^{(k)} - \mathbf{A}\|_F$ is stochastically bounded and does not increase with n. In Appendix A, we empirically show that prominent graphon estimators (E. M. Airoldi et al., 2013; Chan & Airoldi, 2014; Chatterjee, 2015; Zhang et al., 2017) satisfy Assumption 2.

Theorem 1. *(Convergence rate of MMV). Let* $V = \frac{1}{n(n-1)/2}\sum_{1\le i<j\le n} p_{ij}(1-p_{ij})$. *Under Assumptions 1 and 2, for any* $M \in \mathcal{M}$ *and any* $\mathbf{A}$ *generated by Eq.*(2.1) *with parameter* $\mathbf{P}_0$, *as* $n \to \infty$, *we have*

$$MMV_M(\mathbf{A}) - MSE(\mathbf{P}^{\mathbf{A}}_M, \mathbf{P}_0) - V = O_p(\frac{1}{n} \vee \frac{1}{K}),$$

where $\vee$ *denotes the max operation.*

Theorem 1 states the MMV score is a consistent estimate of $\text{MSE}(\mathbf{P}^{\mathbf{A}}_M, \mathbf{P}_0) + V$ with a error rate of $\frac{1}{n} \vee \frac{1}{K}$. When $K = o(n)$, the convergence rate is $\frac{1}{K}$, and larger K leads to faster convergence rate. When $K = \Omega(n)$, the convergence rate is $\frac{1}{n}$, in which case increasing K does not increase the convergence rate.

Remark. Since increasing K comes with an increased computational cost, and $K = O(n)$ is enough to get the optimal convergence rate, we suggest $K = O(n)$ in practice.

Theorem 1 is crucial for justifying the selection consistency property of MMV score. To derive the selection consistency of MMV, we impose the following Assumption 3.

Assumption 3. (Distinguishable rate). *For any* $M^{mse}_c \in \mathcal{M}/\{M^{mse}_o\}$ *and any* $\mathbf{A}$ *generated by Eq.*(2.1) *with parameter* $\mathbf{P}_0 = (p_{ij})$, *we have*

$$MSE(\mathbf{P}^{\mathbf{A}}_{M^{mse}_c}, \mathbf{P}_0) - MSE(\mathbf{P}^{\mathbf{A}}_{M^{mse}_o}, \mathbf{P}_0) = \Omega_p(\frac{1}{n} \vee \frac{1}{K}),$$

where $\Omega_p(\cdot)$ *describes the asymptotic lower bound.*

Assumption 3 essentially states how the MSE is affected by the hyperparameter M^{mse}_c and M^{mse}_o. We refer to the difference of MSE induced by these hyperparameters as the distinguishable rate. As suggested by the aforementioned analysis, we set $K = O(n)$, in which case the distinguishable rate is lower bounded by $\frac{1}{n}$. Assumption 3 is a mild condition since many graphon estimation methods satisfy this assumption. For example, the universal singular value thresholding (USVT), which is a widely used graphon estimation method (Chatterjee, 2015), satisfies Assumption 3. In particular, from Theorem 2.3 of Chatterjee, 2015, we can rewrite the $\text{MSE}(\mathbf{P}^{\mathbf{A}}_M, \mathbf{P}_0)$ as $C(M)a_n$ for any $M \in \mathcal{M}$, where $C(\cdot)$ is a function of the

hyperparameter M (independent of n), and a_n is a random variable lower bounded by $\frac{1}{\sqrt{n}}$. It is easy to verify that $\text{MSE}(\mathbf{P}^{\mathbf{A}}_{M_c^{mse}}, \mathbf{P}_0) - \text{MSE}(\mathbf{P}^{\mathbf{A}}_{M_o^{mse}}, \mathbf{P}_0)$ is lower bounded by $\frac{1}{n}$.

Theorem 2. *(Selection consistency). Under Assumptions 1,2,3, for any $M_c \in \mathcal{M}/\{M_o\}$ and any $\mathbf{A}$ generated by Eq.(2.1) with parameter $\mathbf{P}_0 = (p_{ij})$, we have*

$$\lim_{n\to\infty} \mathbb{P}\left(MSE(\mathbf{P}^{\mathbf{A}}_{M_o}, \mathbf{P}_0) < MSE(\mathbf{P}^{\mathbf{A}}_{M_c}, \mathbf{P}_0)\right) = 1.$$

Theorem 2 is a key conclusion in this paper, providing the selection consistency property of MMV. It states that the optimal hyperparameter selected by MMV, i.e., M_o asymptotically leads to the smallest MSE that is a gold standard to measure the estimation error. This indicates that M_o is asymptotically consistent with M_o^{mse}.

2.4 Implementation of MMV

Recall that in Algorithm 1, we need to specify two parameters: (1) the number of folds K and (2) the probability parameter θ. In this section, we present how to set K and θ when implementing MMV.

Selection of K. To ensure the asymptotic consistency, we set $K = O(n)$, i.e., $K = \lambda n$, where $\lambda \in (0, 1)$. Our numerical robustness experiments suggest (see Appendix A) that the accuracy of MMV is not sensitive to $\lambda \in (0.1, 0.9)$. In this paper, we set $K = 0.1n$ for simplicity.

Selection of θ. Since $\tilde{a}_{ij}^{(k)}$ and a_{ij} follows two different Bernoulli distributions with mean $\mathbb{P}(\tilde{a}_{ij}^{(k)} = 1)$ and p_{ij} respectively, we calculate $(\mathbb{P}(\tilde{a}_{ij}^{(k)} = 1) - p_{ij})^2$ to quantify the difference. By Eq. (2.7), we have

$$(\mathbb{P}(\tilde{a}_{ij}^{(k)} = 1) - p_{ij})^2 = (\frac{\theta - p_{ij}}{K})^2.$$

For $i, j = 1, \ldots, n$, we calculate the sum of differences, i.e.,

$$\sum_{i<j} (\mathbb{P}(\tilde{a}_{ij}^{(k)} = 1) - p_{ij})^2 = \frac{1}{K^2} \sum_{i<j} (\theta - p_{ij})^2. \tag{2.12}$$

An ideal case is that $\tilde{a}_{ij}^{(k)}$ and a_{ij} come from the same distribution for any i, j. In this case, Eq. (2.12) is reduced to zero. Nevertheless, this ideal case is impractical since p_{ij} is unknown. We thus aim to find a θ that minimizes Eq. (2.12). A simple calculation shows that when θ is equal to $\bar{p} = \frac{\sum_{i<j} p_{ij}}{n(n-1)/2}$, Eq. (2.12) is minimized. Since $\bar{p}$ is also unknown, we use $\bar{a} = \frac{\sum_{i<j} a_{ij}}{n(n-1)/2}$ as an estimator of $\bar{p}$. Here $\bar{a}$ measures the network density, i.e., the ratio of observed edges to the number of possible edges for a network. It is easy to show that $\bar{a}$ is a best unbiased estimator of $\bar{p}$.

Analogous to the conventional CV, MMV pretends not to see validation data when training, in order to avoid over-fitting. When calculating the network density, we leave out the information contained in the validation data. We thus set

$$\theta = \frac{\sum_{v_i v_j \in \mathcal{S}_{-k}} a_{ij}}{|\mathcal{S}_{-k}|}. \tag{2.13}$$

Based on the above analysis, setting θ to be the network density using Eq (2.13) is a good choice. In addition, our empirical results in the Appendix A show that the performance of MMV is always competitive when θ is selected to be the network density, which empirically confirms our proposal.

2.5 Simulation Studies

We generate networks using four different graphon functions that are widely used in the literature (Chan & Airoldi, 2014). The detailed graphon functions are listed in Table 2.1. The rank and the density are measured numerically using the average rank of $\mathbf{P}_0$ (200×200 size) over 100 replications. These four graphons all generate a low-rank probability matrix. Graphon 1 and graphon 2 generate dense networks, while Graphon 3 and graphon 4 generate sparse networks. Figure 2.2 visualizes the corresponding heat-map of a 200×200 probability matrix $\mathbf{P}_0$ generated by each graphon function. High values in $\mathbf{P}_0$ are highlighted in red, while low values are colored in blue.

Under each graphon setting, we generate networks of different sizes to investigate the asymptotic performance, i.e., we vary the number of nodes $n \in \{50, 100, 150, 200\}$. All results are based on 100

replications. We compare MMV with ECV (Li et al., 2020). All experiments are implemented on a machine with a 40-core CPU and 192 GB of RAM.

Table 2.1: List of four graphons considered in this simulation study.

Graphon ID	1	2	3	4
Function	$\frac{1}{1+\exp(-10(\mu_i^2+\mu_j^2))}$	$0.5+\frac{\mu_i\mu_j}{2}$	$\mu_i\mu_j$	$\exp(-(\mu_i^{0.7}+\mu_j^{0.7}))$
Rank	74	3	1	1
Network Density	0.95	0.63	0.26	0.33

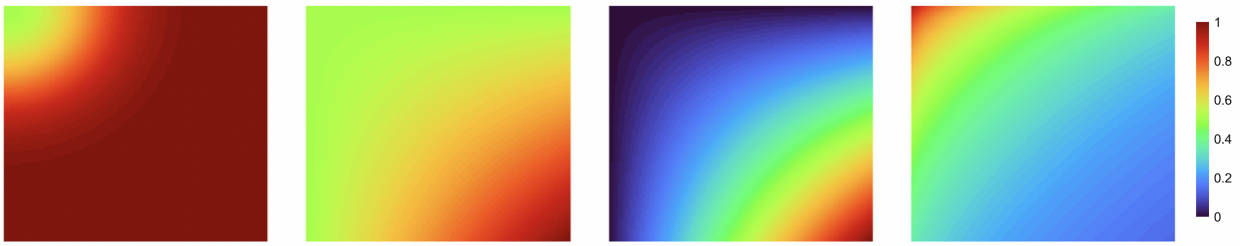


Figure 2.2: Heatmap of a 200×200 $\mathbf{P}_0$ generated by graphons 1-4, from left to right.

2.5.1 Hyperparameter Selection

In this subsection, we demonstrate the performance of MMV in tuning the hyperparameter of two widely-used graphon estimation methods: neighborhood smoothing (NS) (Zhang et al., 2017) and sort-and-smooth (SAS) (Chan & Airoldi, 2014).

Hyperparameter Selection in NS

In particular, NS depends on h which controls the neighborhood size. Zhang et al., 2017 suggests $h = c\sqrt{\frac{\log n}{n}}$, where $c > 0$ is a predefined hyperparameter. NS sets $c = 1$ by default. We consider tuning c from the candidates $\{0.5, 1, \ldots, 5\}$. The selected hyperparameter is the one minimizing the MMV score.

We conducted simulation results to validate our Theorem 2, that is, MMV score in Eq. (2.10) and MSE in Eq. (2.3) are minimized at the same hyperparameter. For visualization purpose, we normalize the MMV score to $(0, 1)$ using $\frac{\text{x}-\text{min}}{\text{max-min}}$. In the same way, we normalize MSE to $(0, 1)$. MMV score

and MSE at each hyperparameter are averaged over 100 replications. Figure 2.3 shows MMV score and MSE with different hyperparameters under $n = 50, 100, 150, 200$ of graphon 1-4. This illustrates that the choice of c can lead to a big difference in MSE, and $c = 1$ is usually not a good choice. In addition, when $n = 50$, the minimizer of MMV and the minimizer of MSE are the same for graphon 1 and graphon 2. When n increases to 200, MMV and MSE have the same minimizer for all graphons. This example illustrates that the choice of c can lead to a big difference in estimation error, and the ECV is successful at choosing it, especially when n is large. These observations empirically confirm our Theorem 2, i.e., selection consistency.

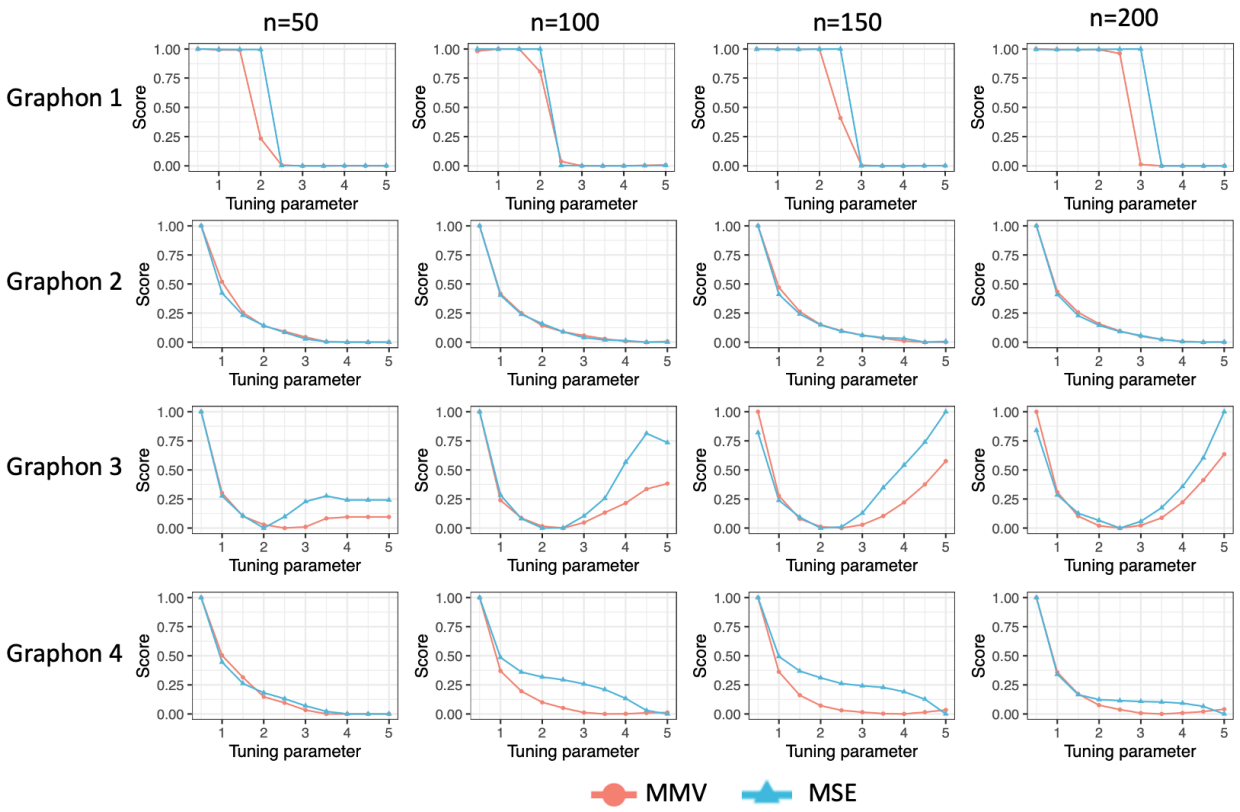


Figure 2.3: MMV score (red) and MSE (blue) with different hyperparameters under $n = 50, 100, 150, 200$ (from left to right) for graphons 1-4 (from top to bottom).

Accuracy. We further measured the estimation error MSE of the estimator using the hyperparameter selected by MMV. We compared it with the estimation error of using ECV, and the estimation error of using the default setting (i.e., $c = 1$). Table 2.2 shows the results of $n = 200$. From Table 2.2, we

observe that the MSE of MMV and ECV are both lower than that of default NS. This indicates that hyperparameter tuning is necessary. Furthermore, we observe that the MSE of MMV is lower than that of ECV. In other words, the hyperparameter selected by MMV yields lower MSE.

Table 2.2: MSE (mean $\pm$ standard deviation) of MMV, ECV and Default NS, all multiplied by 100. MSE is averaged over 100 replications.

Graphon ID	1	2	3	4
MMV	**0.50 $\pm$ 0.06**	**0.79 $\pm$ 0.11**	**0.94 $\pm$ 0.03**	**1.39 $\pm$ 0.11**
ECV	18.29 $\pm$ 20.91	0.83 $\pm$ 0.09	0.99 $\pm$ 0.05	1.42 $\pm$ 0.08
Default NS	39.05 $\pm$ 3.33	1.06 $\pm$ 0.06	1.02 $\pm$ 0.04	1.51 $\pm$ 0.07

Time. Figure 2.4 compares the computational time of hyperparameter tuning of MMV and that of ECV. From Figure 2.4, we have two observations. First, the computational cost of MMV and ECV seem to have the same order. Since NS requires $O(n^3)$ computation, the computational cost of MMV and ECV have the same order, i.e., $O(Kmn^3)$. Second, even though in the same order, the computational time of MMV is less than that of ECV under all settings.

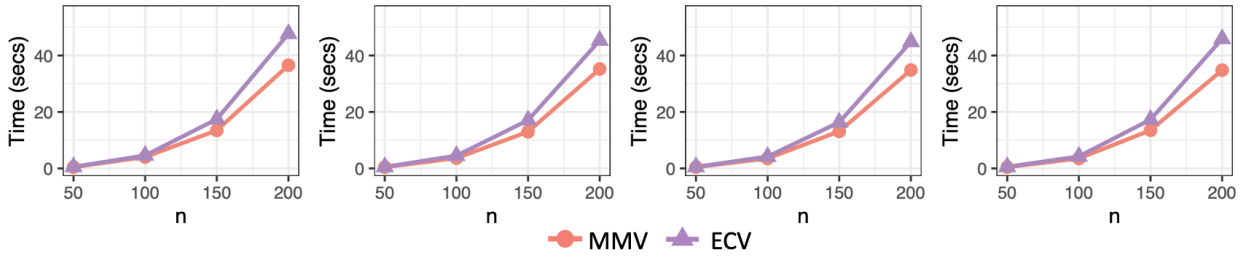


Figure 2.4: The average computational time (in seconds) for graphon 1-4, from left to right, over 100 replications. The red and purple lines represent the MMV and ECV, respectively.

Hyperparameter Selection in SAS

Chan and Airoldi, 2014 proposes a method called sort-and-smooth (SAS) to estimate $\mathbf{P}_0$. SAS depends on a predefined hyperparameter, i.e., the number of blocks b. SAS set $b = 2$ by default. We consider using MMV to select the optimal b from candidates $\{2, 4, \ldots, 30\}$. Figure 2.5 shows MMV score and MSE with different hyperparameters under $n = 50, 100, 150, 200$ (from left to right) of graphon 1-4 (from

top to bottom). As we can see, when $n = 50$, the minimizer of MMV and the minimizer of MSE are the same for graphon 4. When n increases to 200, MMV and MSE have the same minimizer for all graphons. This empirically confirms Theorem 2.

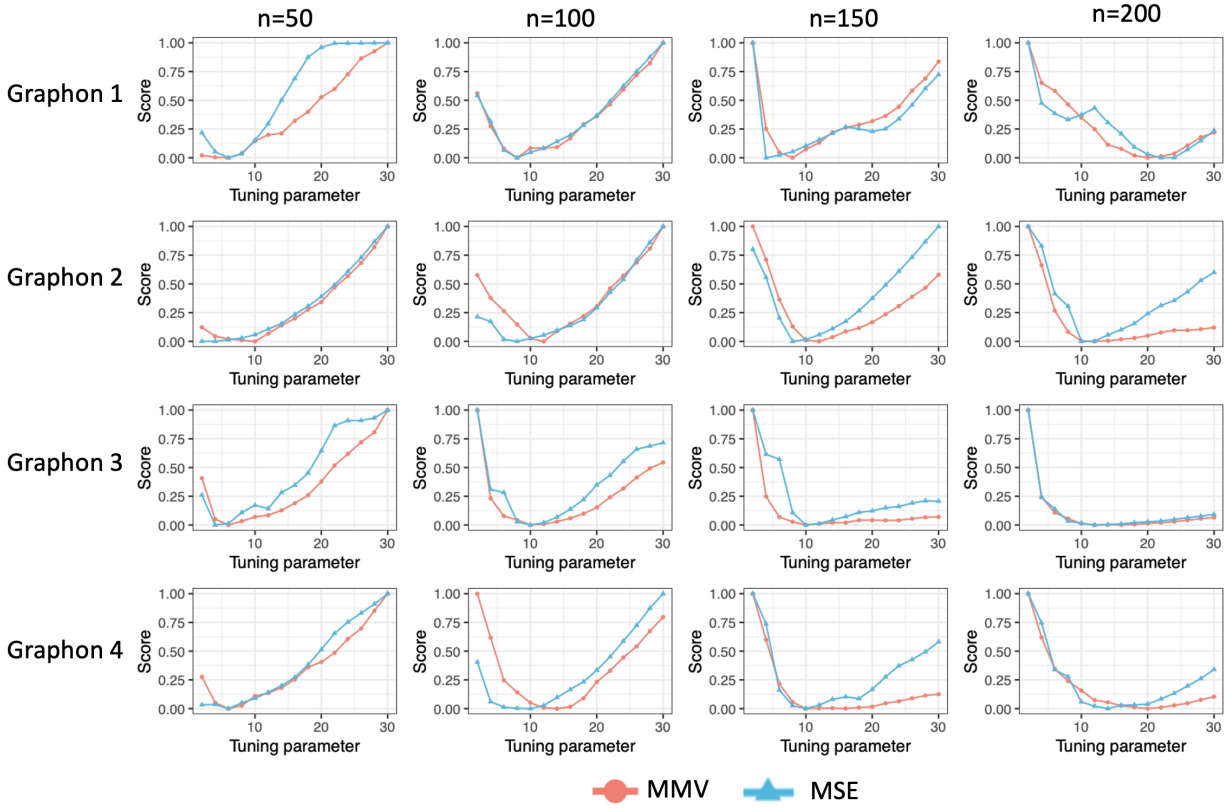


Figure 2.5: MMV score (red) and MSE (blue) with different hyperparameters under $n = 50, 100, 150, 200$ (from left to right) for graphons 1-6 (from top to bottom).

Accuracy. We further measured the estimation error MSE of the estimator using the hyperparameter selected by MMV. We compared it with the estimation error of using ECV, and the estimation error of using the default setting (i.e., $b = 2$). Table 2.3 shows the results of $n = 200$. From Table 2.3, we observe that the MSE of MMV is lower than that of ECV under all settings.

Time. Figure 2.6 compares the computational time of hyperparameter tuning of MMV and that of ECV. From Figure 2.6, we have two observations. First, the computational cost of MMV and ECV seem to have the same order. Since NS requires $O(n^3)$ computation, the computational cost of MMV and

Table 2.3: MSE (mean $\pm$ standard deviation) of MMV, ECV and Default SAS, all multiplied by 100. MSE is averaged over 100 replications.

Graphon ID	1	2	3	4
MMV	**0.51 $\pm$ 0.13**	**1.12 $\pm$ 0.24**	**1.16 $\pm$ 0.09**	**1.42 $\pm$ 0.07**
ECV	0.64 $\pm$ 0.26	1.21 $\pm$ 0.22	1.21 $\pm$ 0.07	1.46 $\pm$ 0.08
Default SAS	0.98 $\pm$ 0.35	1.66 $\pm$ 0.15	4.21 $\pm$ 1.31	2.22 $\pm$ 0.22

ECV have the same order, i.e., $O(Kmn^3)$. Second, even though in the same order, the computational time of MMV is less than that of ECV under all settings.

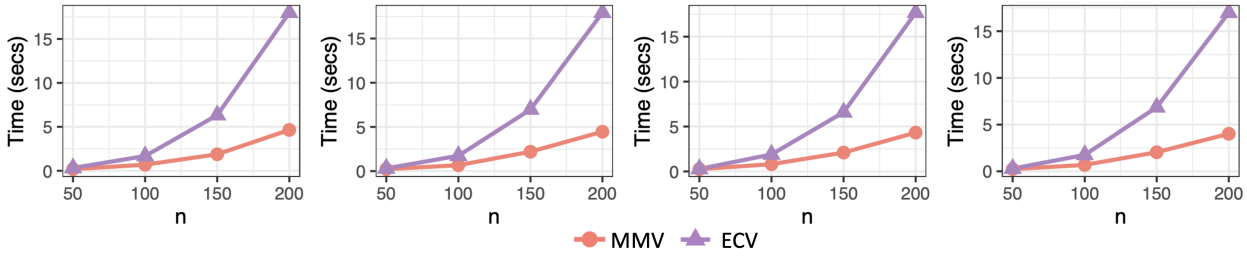


Figure 2.6: The average computational time (in seconds) for graphon 1-4, from left to right, over 100 replications. The red and purple lines represent the MMV and ECV, respectively.

2.5.2 Method Comparison

Even though various graphon estimation methods have been proposed, there is no uniformly best method that can beat all other methods for all data. Given a particular data, it is usually unknown which graphon estimation method we should apply. Given m different graphon estimation methods, the optimal one should lead to the smallest MSE, which is impractical to calculate in real-world applications, since MSE requires knowing the true probability matrix.

Fortunately, our developed MMV does not require knowing the true probability matrix. Thus, besides hyperparameter selection, MMV can also be used as a criterion for method comparison. In particular, for each candidate method, we can get an estimator which is further used to calculate MMV score. A smaller MMV score indicates a smaller MSE, supported by Theorem 2. We select the method that leads

to the smallest MMV score. Actually, the true optimal method is defined as the one that leads to the smallest MSE. We calculate the fraction of times the true optimal method is correctly selected among 100 replications. This fraction value is called accuracy.

We consider five widely used graphon estimation methods as our candidates, including (1) M_1: stochastic block model approximation (SBA) algorithm (E. M. Airoldi et al., 2013) which requires block structure assumption, (2) M_2: sort-and-smooth (SAS) method (Chan & Airoldi, 2014) which requires a monotone degree assumption, (3) M_3: universal singular value thresholding (USVT) algorithm (Chatterjee, 2015) which requires the low-rank assumption, (4) M_4: neighborhood smoothing (NS) method (Zhang et al., 2017), and a most recent method (5) M_5: iterative connecting probability estimation method (ICE) (Qin et al., 2021).

Accuracy. The accuracy results are shown in the first row of Figure 2.7, from which we have the following observations. First, the accuracy of MMV increases with network size. This behavior confirms the selection consistency property stated in Theorem 2. Second, across all four graphon settings, MMV performs the best. When $n = 200$, MMV can select the optimal method with 100% accuracy for all settings. **Time.** The second row in Figure 2.7 compares the time of MMV with that of ECV. The computational time of MMV is smaller than ECV under all settings. All these observations suggest that MMV is an efficient tool.

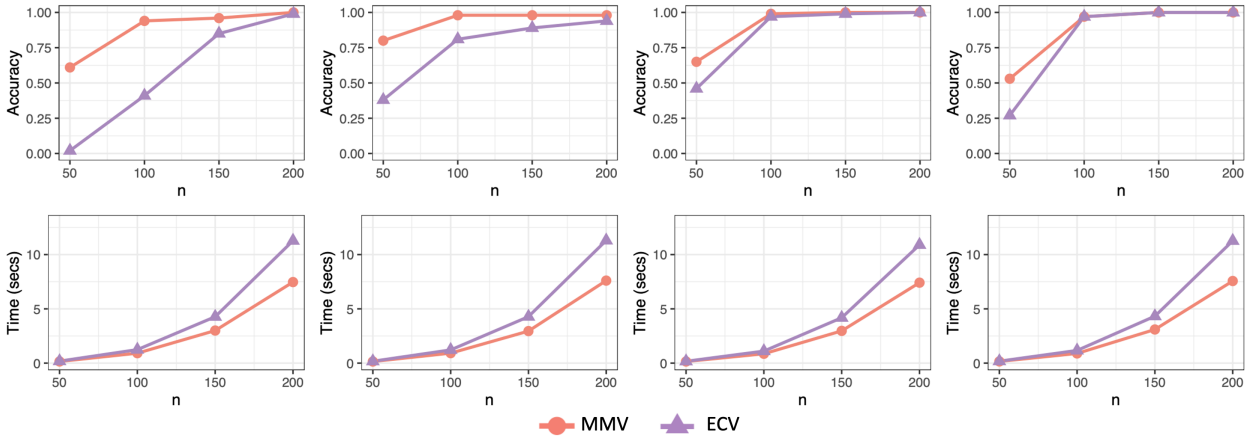


Figure 2.7: The top row shows the accuracy for graphon 1-4 (from left to right). The bottom row shows the average computational time (in seconds) for graphon 1-4 (from left to right), over 100 replications. The red and purple lines represent the MMV and ECV, respectively.

2.6 Real-world Applications

For real networks, since the true probability matrix $\mathbf{P}_0$ is unknown, it is infeasible to calculate the MSE. Thus the evaluation metrics used in the simulation studies are inapplicable in real-world applications. We thus assess the practical utility of MMV by applying it to downstream analysis, i.e., link prediction. Particularly, we use the accuracy of link prediction to evaluate the performance of MMV in tuning hyper-parameters. We consider link prediction in two scenarios. In the first scenario, we consider that network evolves over time, and we predict new links that are likely to appear in the near future. Results are shown in Section 2.6.1. In the second scenario, we consider that network is static, and we predict missing links that are not observed. Results are shown in Section 2.6.2.

2.6.1 Drug Repurposing through Disease-Drug Association Prediction

Drug repurposing identifies novel uses for existing drugs, and has become a promising strategy, due to the notorious advantages over traditional drug discovery (Gysi et al., 2021). Traditional drug discovery is a tremendously time-consuming, costly, and difficult task. Generally, the process takes 17 years and costs an average of US $2.6 billion. Furthermore, only approximately 2.01% of all drug development candidates finally make it to the market as a successful treatment. During past years, successful repurposing examples include: Minoxidil, originally approved to treat hypertension, has been repurposed to treat hair loss (Santamaria et al., 2021); Remdesivir, originally developed for treating ebola, has demonstrated effectiveness of treating COVID-19 (Al-Tawfiq et al., 2020); Hydroxychloroquine, whose original indication is malaria, has been used for treating COVID-19 (Liu et al., 2020).

Constructing a disease-drug network encoding disease–drug relationship is key to drug repurposing. Many unidentified relationships are buried in the sheer volume of biomedical literature. To extract the relationship in the vast literature, we employ a straightforward approach, that is, constructing a disease-drug co-occurrence network where a node represents a drug or a disease, and two nodes are connected if

they both occurred in a PubMed publication. Herein, we focus on the abstract of each publication, since the abstract of a biomedical article typically has the direct expression of the author's research conclusions.

In this paper, we collected the abstracts of 5496 PubMed literature related to COVID-19 between 01/01/2020 and 04/30/2020. We then extracted disease and drug entities using a well-known entity annotation tool PubTator (Wei et al., 2013). We further drew an edge between a drug and a disease if they both occurred in the same publication. There are 280 nodes and 952 edges in the constructed network, and the network density is only 0.02. Our constructed network is visualized in Figure 2.8, from which we can see, two central nodes are COVID-19 and hydroxychloroquine.

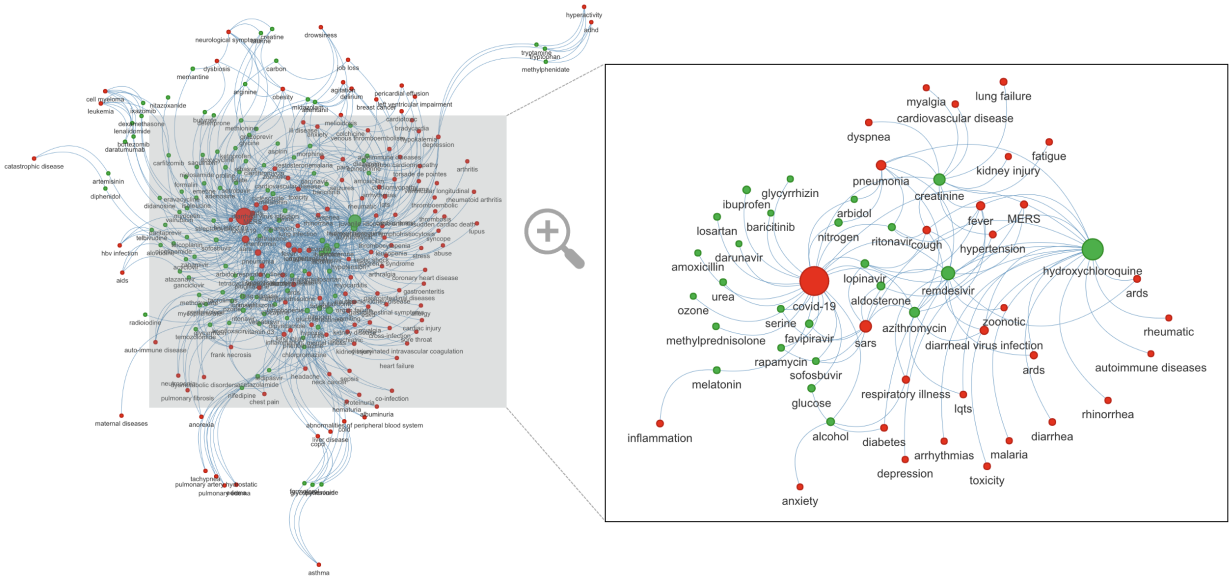


Figure 2.8: Disease-drug co-occurrence network, where red and green nodes represent disease and drugs, respectively. We zoom in on this network to get a subgraph, which is visualized on the right. The size of a node is proportional to the node degree.

For this network, we applied the NS method (Zhang et al., 2017) to get the estimator. Recall that the hyperparameter in NS is c. Figure 2.9 shows MMV score and ECV score under different tuning hyperparameters, i.e., $c \in \{0.2, 0.4, \ldots, 2\}$. As we can see, MMV is minimized at $c = 1.2$, while ECV is minimized at $c = 0.4$. By setting $c = 1.2$ or $c = 0.4$, we get two estimators of the probability matrix. The ijth entry in the estimated probability matrix indicates the probability of nodes i and j having an association. Thus, if a disease-drug pair has no edge yet, and has a high probability, we predict this disease-

drug pair might have an association in the future. Based on decreasing order of the estimated probability for disease-drug pairs that are not connected yet, we output top q most likely future disease-drug links, where q will be considered at different levels.

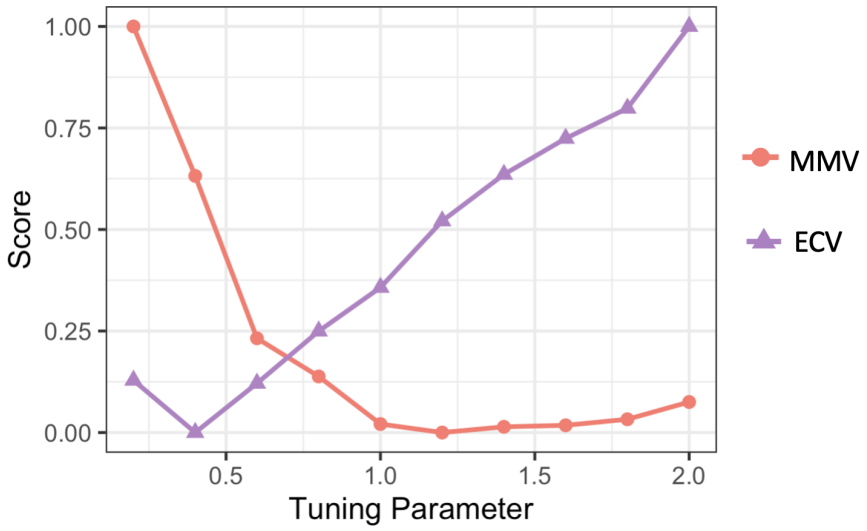


Figure 2.9: MMV score and ECV score with different hyperparameter.

We further conduct follow-up confirmatory analysis to investigate whether these predicted links are correct. Specifically, for each predicted link involving a drug and a disease, we investigate whether they co-occur in the "future" PubMed literature. Since we constructed the disease-drug network only using the information contained in the literature published between 01/01/2020 and 04/30/2020, we define "future" PubMed literature as the literature published after 04/30/2020. Since there are almost 300,000 COVID-19 publications after 04/30/2020, scanning all those publications requires a high computational cost. We thus only consider short-term "future" literature, i.e., literature published between 05/01/2020 and 05/15/2020. If the predicted disease-drug pair occur in future publications, we conclude this predicted link is correct. Herein, we define accuracy as the proportion of disease-drug pairs that occur in future publications among q predictions.

We consider $q = 10, 20, \ldots, 100$. The accuracy under different q of MMV and ECV is shown in Figure 2.10. In general, the accuracy decreases as q increases, since larger q means including more disease-drug pairs with lower probability, which naturally leads to inaccuracy. The accuracy of MMV and ECV

both reach 100% when $q = 10$ and $q = 20$. When q is greater than 30, the accuracy of MMV is greater than that of ECV. Thus, MMV yields more accurate predictions. In addition, the computational time of MMV is 56.76 Seconds, while that of ECV is 71.82 Seconds.

In addition, we further investigate the top three predicted disease–drug pairs using MMV, i.e., {diarrheal, creatinine}, {zoonotic, hydroxychloroquine} and {COVID-19, ledipasvir}. Ledipasvir is a drug developed for the treatment of the hepatitis C virus, our results suggest that ledipasvir might be associated with COVID-19. An interesting fact is that more and more research confirms our finding, showing that ledipasvir in combination with sofosbuvir has great potential to inhibit SARS-CoV-2 replication (Pirzada et al., 2021). Actually, ledipasvir has completed phase 3 trials for COVID-19 Treatment.

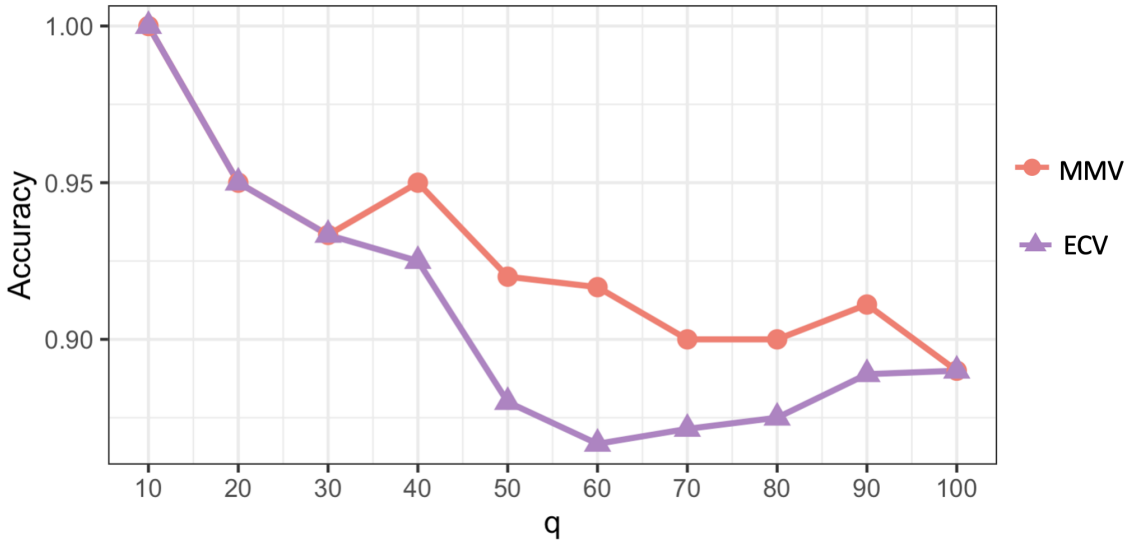


Figure 2.10: Accuracy of MMV and ECV for different q.

2.6.2 Link Prediction for a Partially Observed Network

Link prediction for a partially observed network assumes that the network is not fully observed and there are some missing links, and the task of link prediction is to find out these links. We considered three widely-used large-scale real-world networks for the link prediction task, including (1) political blogs network (1222 nodes and 16714 edges) (Adamic & Glance, 2005), (2) coauthorships in network science (1589 nodes and

2742 edges) (Newman, 2006), (3) yeast protein-protein interaction network (2617 nodes and 11855 edges) (Von Mering et al., 2002).

Following Zhang et al., 2017, we conduct a semi-supervised procedure to evaluate the performance of link prediction. Let the $\mathbf{A}^0 = (a_{ij}^0)$ denote the adjacency matrix of the real-world network. We first obtain a new adjacency matrix $\mathbf{A}^1(a_{ij}^1)$ with 10% edges randomly removed from $\mathbf{A}^0$, i.e., $a_{ij}^1 = b_{ij}a_{ij}^0$, where $b_{ij} \sim \text{Ber}(0.9)$. We then obtained an estimator using $\mathbf{A}^1$, where MMV or ECV is applied for tuning the hyperparameter. Since the aforementioned NS requires $O(n^3)$ computation, implementing NS for networks with thousands of nodes such as coauthorships in network science is very expensive. Here we consider tuning the hyperparameter in a fast graphon estimation method, that is, SAS (Chan & Airoldi, 2014). SAS needs a predefined parameter: the number of blocks. We applied MMV or ECV to select an appropriate number of blocks from $\{10, 20, \ldots, 100\}$. Finally, we assessed the performance of link prediction using the area under curve (AUC) by comparing a_{ij}^0 with the estimated probability for node pairs with $b_{ij} = 0$.

Table 2.4 shows an average AUC of 100 replications. From Table 2.4, MMV outperforms ECV for dataset PolBlog and NetSci, and achieves competitive performance for Yeast. Table 2.4 also shows the average time (in minutes). The computational cost of MMV is largely reduced by MMV. Since SAS only requires $O(n^2)$ computation, the computational cost of MMV is $O(n^2)$, while that of ECV is $O(n^3)$. This further validates the computational advantage of MMV.

Table 2.4: AUC (average $\pm$ standard deviation) and computational time in minutes (average $\pm$ standard deviation) of MMV and ECV, over 100 replications.

		PolBlog	NetSci	Yeast
AUC	MMV	**0.88 $\pm$ 0.01**	**0.72 $\pm$ 0.01**	0.80 $\pm$ 0.02
AUC	ECV	0.80 $\pm$ 0.02	0.70 $\pm$ 0.01	0.80 $\pm$ 0.02
Time (minutes)	MMV	**56.90 $\pm$ 0.13**	**51.01 $\pm$ 3.96**	**240.90 $\pm$ 16.22**
Time (minutes)	ECV	258.65 $\pm$ 2.11	771.23 $\pm$ 10.00	6021.12 $\pm$ 18.72

2.7 Conclusions and Discussions

The contribution of our MMV procedure to the hyperparameter selection in graphon estimation is four-fold. First, it does not impose any assumption on the graphon function form. Therefore, MMV can be considered a model-free procedure applicable in any random network model. Second, it has the theoretical guarantee of selecting the optimal hyperparameter. Third, the implementation of MMV is computationally effective, user-friendly, and essentially tuning-free. Fourth, as demonstrated by our simulation studies, MMV has superior performance. We believe that the MMV should become an indispensable member of the repository of validation tools and we recommend its broad use.

Future extensions. The general scheme of creating a masked mirror "image" of the adjacency matrix is useful for other bootstrapping-based methods. An interesting future direction we plan to investigate is the conformal inference for graphon by creating mirror "images". Another future direction is to develop MMV with a confidence interval that takes into account the randomness in the MMV score. Instead of outputting which candidate is the best, MMV with confidence interval outputs a confidence set with guaranteed probability under certain regularity conditions. In addition, it is also interesting to extend MMV to a multi-layer network, which is increasingly common. We leave these interesting directions for future work.

Chapter 3

Network Sampling Using Ricci Curvature

3.1 Introduction

As we enter the big data era, our capacity to access large graphs (a.k.a. networks) has provided unprecedented opportunities and challenges. For example, T. Wang et al., 2011 presented a Twitter social network, which has more than 190 million nodes (users) who generate more than 65 million edges (tweets) every day (T. Wang et al., 2011). Such huge networks enable researchers to tackle more complex problems. However, they pose great challenges to storing, visualizing, and analyzing since their sheer volumes render many computational methods infeasible.

Graph subsampling. Graph subsampling is a commonly used technique to address this scalability issue because of its simplicity and efficiency. Graph subsampling aims to take a subgraph that preserves critical features of the full graph. Various graph subsampling methods that preserve different graph properties have been proposed, including node sampling (Mall et al., 2013; Zeng et al., 2019), edge sampling (Krishnamurthy et al., 2005), and exploration sampling (Goodman, 1961; Hübler et al., 2008; Leskovec et al., 2005; Maiya & Berger-Wolf, 2010). Researchers evaluate the graph subsampling approaches by measuring the similarity between the features of the original graph and those of the subgraph. The features

include, e.g., degree distribution (Adamic et al., 2001), minimum cut (Hu & Lau, 2013), and the number of triangles (Seshadhri et al., 2014).

An important graph feature that has received less attention is the number of communities (denoted by M), which plays a crucial role in identifying the community structures. Network data often have natural communities, and the identification of these communities helps answer vital questions in a variety of fields (Rohe et al., 2011). For example, communities in social networks may represent groups of people who share a similar interest, and communities in protein-protein interaction networks could be regulatory modules of interacting proteins (Rohe et al., 2011). In this paper, we focus on the setting where M is a fixed model parameter, and there is a ground truth about M. This setting is widely used in many models, e.g., stochastic block model (SBM) (Holland et al., 1983) and its variants such as degree-corrected SBM (DCBM) (Karrer & Newman, 2011). The SBM family are arguably the most widely-used generative model for community detection from a theoretical perspective. In fact, Vaca-Ramırez and Peixoto, 2022 performed a systematic analysis of the quality of fit of the SBM for 275 real networks, and observed that "SBM is capable of providing an accurate description for the majority of networks considered". Indeed, there are other settings where the number of communities is not fixed. For example, Olhede and Wolfe, 2014 used SBM to approximate a nonparametric graphon model, under which case the number of communities is a hyperparameter and is not fixed. The latter hyperparameter case is beyond the scope of this paper.

Under the setting where M is a model parameter, many community detection methods have been proposed, such as modularity maximization (Good et al., 2010; Newman, 2006), spectral clustering (Liu et al., 2018; Rohe et al., 2011; Von Luxburg, 2007), and pseudo-likelihood based methods (Amini et al., 2013; J. Wang et al., 2021). The theoretical properties of most community detection methods, such as consistency and asymptotic distributions, are built based on the assumption that M is known (Ma et al., 2021). In addition, M is usually required as an input for those community detection algorithms. However, in practice, we do not have the information of M, which significantly diminishes the usefulness of the aforementioned methods. Existing methods for estimating M is usually very expensive. For example, the cross-validation method proposed by Li et al., 2020 requires a computational cost that is cubic

in the number of nodes n. When there are thousands or millions of nodes, the computational cost is unaffordable.

Thus, it is highly desirable for subsampling methods to yield subgraphs with $\tilde{n} << n$ nodes preserving the number of communities M, such that we can use it to get an accurate estimation while reducing the computational cost. Despite many successful applications, existing subsampling methods tend to leave out minority communities (i.e., a community with a smaller number of nodes) because nodes with high degrees are more likely to be sampled into subgraphs. Consequently, these subsampling methods usually underestimate M, especially for graphs with imbalanced community structures.

To overcome the shortcomings of existing methods, we develop a graph subsampling method that yields subgraphs that can be used to accurately estimate M. Achieving this goal is challenging since the community structure is hidden and unavailable. Fortunately, recent studies indicate that the community structure is a geometric phenomenon by considering a graph as a Riemannian geometric object (Ni et al., 2015). Some insights into the community information can be obtained by applying some geometric methods to a graph (Ni et al., 2015; Ni et al., 2019; Sia et al., 2019).

Ollivier Ricci Curvature of Graph. In particular, a graph can be regarded as a discrete version of a Riemannian manifold (Ni et al., 2019). A node of the graph is analogous to a point on a Riemannian manifold, and a pair of connected nodes in a graph is analogous to two points connected by a geodesic on a manifold. The partition of a graph into communities is analogous to the geometric decomposition of a Riemannian manifold (Ni et al., 2019). Ricci curvature is a key tool for the geometric decomposition of a Riemannian manifold. It measures how the Riemannian manifold deviates from the flat manifold. Recently, Lin et al., 2011 defines an analog of Ricci curvature for the graph, i.e., the Ollivier Ricci (abbreviated as OR) curvature. Previous empirical results have shown the OR curvature is related to the connectivity of the graphs (Gosztolai & Arnaudon, 2021; Ni et al., 2015). However, the relationship between OR curvature and connectivity is insufficiently explored in theoretical evidence. In this paper, we theoretically show that the OR curvatures of edges within a densely connected community are asymptotically larger than those of edges between two sparsely connected communities, as the size of the graph increases.

Ollivier Ricci Curvature Gradient Based Graph Subsampling. Based on our theoretical result, we propose an OR curvature gradient-based graph subsampling algorithm (abbreviated as ORG-sub). Specifically, ORG-sub randomly chooses one edge as the starting point of the subgraph and calculates the OR curvature of the selected edge. ORG-sub then gradually expands the subgraph by taking the next edge whose OR curvature shows the largest difference from the OR curvature of the previously taken edge. Here, we define the difference as the OR curvature gradient (ORG). We use the ORG to guide the expansion of ORG-sub, i.e., direct ORG-sub which edge to take next. All edges that ORG-sub took expand into the final subgraph. Our proposed ORG-sub enjoys two main advantages. First, ORG-sub has a rigorous theoretical guarantee. In particular, under the SBM scenario, we prove that the probability of ORG-sub taking all communities into the final subgraph converges to one faster than the random walk algorithm, indicating that even nodes in the minority community are subsampled. More importantly, we theoretically show that the estimation of M by the subsampled graph converges to M of the full graph. Second, our extensive empirical experiments based on the simulated and real-world datasets show that the estimator based on ORG-sub subgraphs accurately estimates M while greatly reducing the computation cost.

3.2 Preliminaries

3.2.1 Notations and Definitions

A graph is made up of nodes that are connected by edges. Considering that directed graphs don't approximate Riemannian manifolds in discrete form since geodesics are always bidirectional, we only focus on undirected graphs in this paper. Denote a graph as $G = \langle V, E \rangle$, where V is the node set and E is the edge set. Operator $|\cdot|$ calculates the cardinality of the set. We then denote the cardinality of the node set in a graph by $|V|$ ($|V| = n$), the number of edges in a graph by $|E|$, and the neighborhood set of a node v by $\delta(v) = \{w|(v, w) \in E\}$. The degree of a node v is defined as the cardinality of its neighborhood: $d_v = |\delta(v)|$. For a subset of V, $S \subseteq V$, the subgraph induced by the subset S is denoted

by $G[S] = (S, E_S)$, where the edge set $E_S = \{(v_S, w_S) \in E | v_S \in S, w_S \in S\}$. The neighborhood of the node set S is defined by $N(S) = \bigcup_{v \in S} \delta(v)$, and the neighboring edge set of edge $e = (u, v)$ is $\Delta(e) = \{(x, y) | x \in \{u, v\}, y \in \delta(x) \backslash \{u, v\}\}$. Let $\hat{M}(G[S])$ denote the estimator of M, obtained by using the subgraph $G[S]$ for estimation.

3.2.2 Riemannian manifold and Ricci curvature

The definition of a Riemannian manifold dates back only to 1930s as it was not really until Whitney's work in 1936 that mathematicians obtained a clear understanding of what abstract manifolds were other than just being submanifolds of Euclidean space. An n-dimensional Riemannian manifold consists of a C^∞ manifold $\mathcal{M}$ and a Riemannian metric g, which is a family of smoothly varying inner products $g_x : \mathcal{T}_x\mathcal{M} \times \mathcal{T}_x\mathcal{M} \to \mathbb{R}$ on the n-dimensional tangent spaces $\mathcal{T}_x\mathcal{M}$ at each point $x \in \mathcal{M}$ (Figalli & Villani, 2011). The Riemannian metric g makes it possible to define geodesic distances, and curvatures. Curvature measures how curved a space is, i.e., quantifies the deviation of a space from being locally Euclidean (Saucan, 2015). Spaces of positive curvature have a local geometry similar to spheres, while those of negative curvature are similar to hyperbolic spaces. There are two important notions of curvature: (i) Sectional curvature, associated to tangent planes; (ii) Ricci curvature, associated to tangent vectors (Saucan et al., 2017).

We first present the definition of sectional curvature in Riemannian geometry. Figure 3.1(a) gives an example of a Riemannian manifold $\mathcal{M}$, in which x is a point of $\mathcal{M}$, v be a a unit tangent vector in the tangent space $\mathcal{T}_x\mathcal{M}$ at x, where $||v||_g = \sqrt{g_x(v, v)} = 1$. There exits exactly one geodesic $\text{geo}_x v$ starting at x with tangent vector v. In Figure 3.1(a), $\text{geo}_x v$ is highlighetd in purple. Starting from the point x and following the geodesic $\text{geo}_x v$ for δ unit time, where $\delta > 0$ and $\delta \to 0$, we obtain the endpoint $y = \exp_x \delta v$. Here, $\exp_x : \mathcal{T}_x\mathcal{M} \to \mathcal{M}$ is known as exponential map which is a function mapping the tangent space to the manifold. The Riemannian distance between x and y on manifold $\mathcal{M}$ is $dist_g(x, y) = \delta ||v||_g = \delta$.

Let $w \in \mathcal{T}_x\mathcal{M}$ be another unit tangent vector at x. For simplicity, we will assume that w is orthogonal to v. Let $\varepsilon > 0$ and $\varepsilon \to 0$, the endpoint of εw starting from x is $\exp_x \varepsilon w$. At point y, there exists a

particular tangent vector w' which would be "the same as" w. We obtain w' by parallel transport of w from x to y along δv. Figure 3.1(b) illustrates the idea of parallel transport. Imagine a man stands on the point x and holds a spear, which represents the vector εw. To transport the vector εw at point x to the point y, this man travels along the geodesic $\text{geo}_x v$ while keeping the spear steady. When the man arrives at the point y, the spear represents the vector $\varepsilon w'$. Starting from x along εw, the endpoint is denoted as $\exp_x \varepsilon w$. Analogously, the endpoint of starting from y along $\varepsilon w'$ is denoted as $\exp_y \varepsilon w'$.

Now we are interested in the Riemannian distance between the $\exp_x \varepsilon w$ and $\exp_y \varepsilon w'$. If $\mathcal{M}$ is a Euclidean space, we will simply get a rectangle, as shown in Figure 3.1(c), thus we have $dist_g(\exp_x \varepsilon w, \exp_y \varepsilon w') = dist_g(x, y) = \delta$. If $\mathcal{M}$ is not a Euclidean space, the difference between $dist_g(\exp_x \varepsilon w, \exp_y \varepsilon w')$ and $dist_g(x, y)$ reflects how curved the space is. The larger the difference is, the more curved the space is. Mathematically, when $\delta \to 0$ and $\varepsilon \to 0$, we have

$$dist_g(\exp_x \varepsilon w, \exp_y \varepsilon w') = \delta \left(1 - \frac{\varepsilon k_x(v, w)}{2} + O(\varepsilon^3 + \varepsilon^2 \delta)\right), \tag{3.1}$$

where $k_x(v, w)$ is the sectional curvature associated to the tangent plane (v, w) at point x, δ is the value of $dist_g(x, y)$.

Second, we present the definition of Ricci curvature, which only depends on one tangent vector v. Ricci curvature along tangent vector v at point x, denoted by $Ric_x(v)$, is obtained by summing $k_x(v, w)$ over all the 2-planes containing v. In an $n-$dimensional Riemannian manifold, the number of 2-planes containing v is n. Thus $Ric_x(v) = \sum_{i=1}^{n} K_x(v, w_i)$, where $(w_1, \ldots, w_n)$ are the basis of the tangent space at point x. Let S_x be the set of endpoints $\exp_x \varepsilon w_i, i = 1, \ldots, n$, S_y be the set of endpoints $\exp_y \varepsilon w_i', i = 1, \ldots, n$, of tangent vectors $\varepsilon w_i'$, where w_i' is obtained by parallel transport of w_i from x to y. As shown in Figure 3.1(d), the S_x and S_y can be viewed as balls with radius ε, centered at x and y, respectively (Ollivier et al., 2010). If we map S_x to S_y using parallel transport, i.e., the point $\exp_x \varepsilon w_i$ is mapped to the point $\exp_y \varepsilon w_i'$. The transportation distance between S_x and S_y, denoted as $Trans(S_x, S_y)$, is the average distance between a point in S_x and its mapped point in S_y, i.e., $Trans(S_x, S_y) = \sum_{i=1}^{n} dist_g(\exp_x \varepsilon w_i, \exp_y \varepsilon w_i'), i = 1, \ldots, n$. Some simple algebra of Equation (3.1)

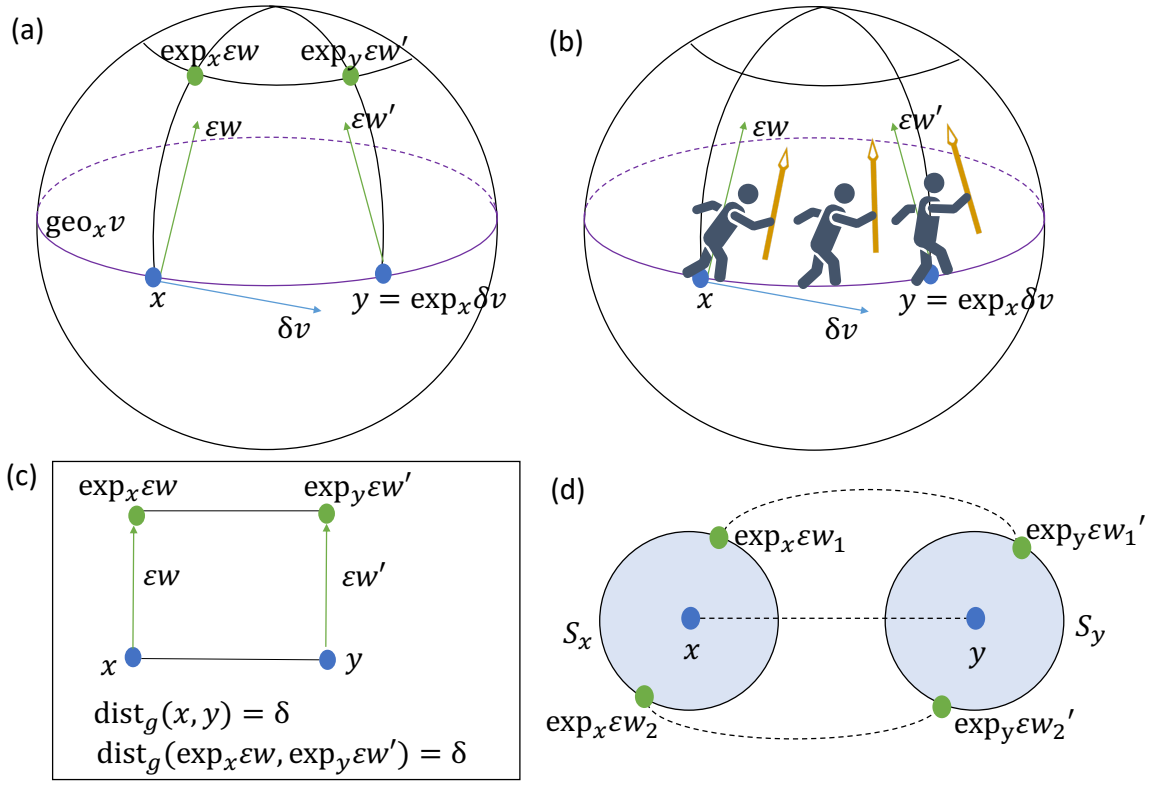


Figure 3.1: (a) The geometric interpretation of sectional curvature. The purple curve represents the $geo_x v$. The blue dots represent x and y, while the green dots represent $\exp_x \varepsilon w$ and $\exp_y \varepsilon w'$. (b) Illustration of parallel transport of tangent vector εw from x to y along δv. (c) Example of flat space. (d) The geometric interpretation of Ricci curvature.

yields

$$Trans(S_x, S_y) = \delta \left(1 - \frac{\varepsilon \sum_{i=1}^{n} k_x(v, w_i)}{2n} + O(\varepsilon^3 + \varepsilon^2 \delta)\right), \tag{3.2}$$

$$= \delta \left(1 - \frac{\varepsilon Ric_x(v)}{2n} + O(\varepsilon^3 + \varepsilon^2 \delta)\right). \tag{3.3}$$

Here, $Trans(S_x, S_y)$ is the distance between two balls S_x and S_y, δ is the distance between the ball centers of S_x and S_y. Intuitively, Ricci curvature measures whether the distance between small balls is

smaller or larger than the distance between the centers of the balls, e.g., when the balls are closer than their centers are, Ricci curvature is positive.

3.2.3 Network and Ollivier-Ricci curvature

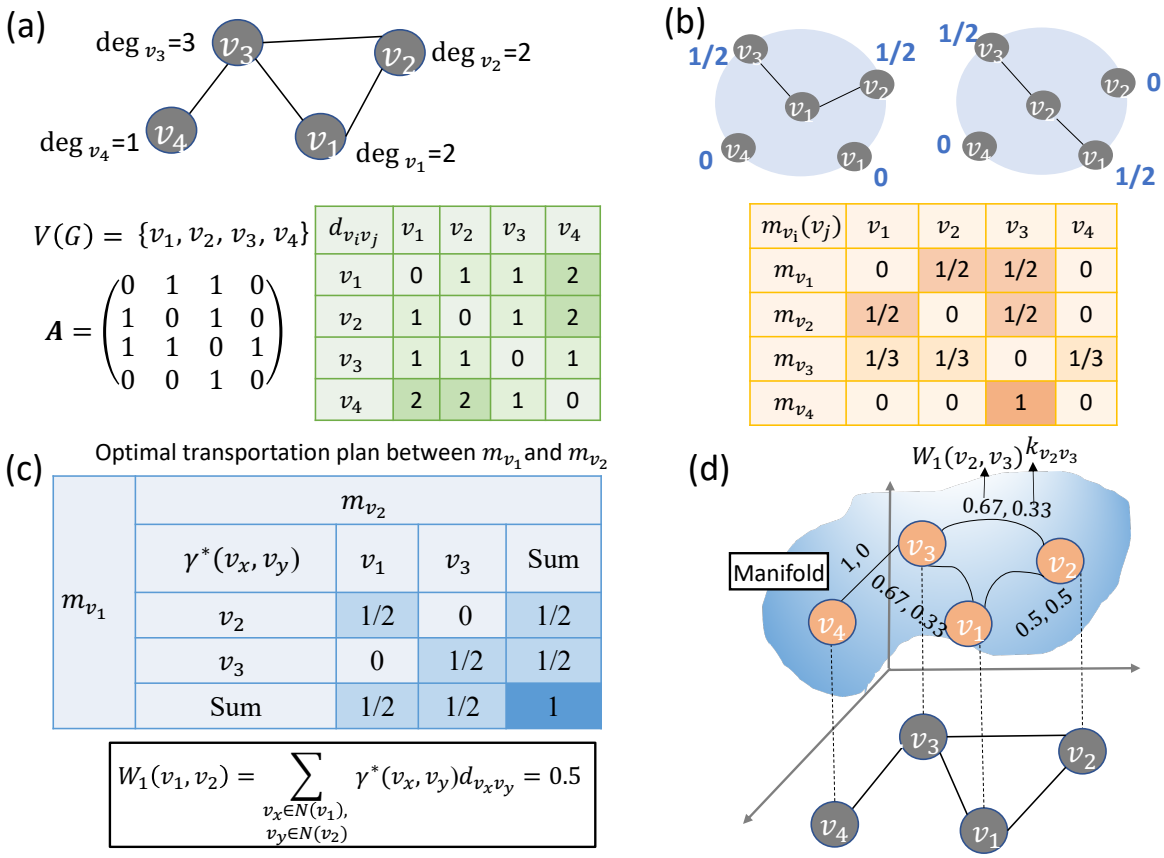
A network can be considered as a discrete approximation to a manifold; on the other hand, a manifold can be considered as a continuous approximation to a network (Zhou & Burges, 2008). The Ricci curvature for network is first introduced by (Ollivier, 2009), thus is named Ollivier-Ricci (OR) curvature.

From Equation (3.3), we know that the calculation of Ricci curvature depends on (i) $Trans(S_x, S_y)$, the distance between balls; (ii) $\delta = dist_g(x, y)$, the distance between the ball centers. To have an analogous definition of Ricci curvature for network, we need to first have an analogous definition of "ball". For each node v_i, we define a probability measure m_{v_i}, which is analogous to a ball centered at node v_i.

We now introduce the definition of OR curvature of the graph. For $\alpha \in [0, 1]$ and any node u with degree d_u, we first define the probability distribution of node u by m_u^α:

$$m_u^\alpha(x) = \begin{cases} \alpha & \text{if } x = u \\ (1-\alpha)/d_u & \text{if } x \in \delta(u) \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

Here α is to keep the probability mass of α at node u itself and distribute the rest uniformly over the neighborhood. Following Ni et al., 2018; Ye et al., 2019, we set $\alpha = 0.5$. This means that each node keeps 50% of the probability mass to itself. Let $d(u, v)$ be the geodesic distance between nodes u and v, which is the shortest path between two nodes of a graph. Figure 3.2(a) shows a network, in which there are four nodes v_1, v_2, v_3, v_4. The degree of the four nodes are $2, 2, 3, 1$, respectively. The green matrix in Figure 3.2(a) shows the geodesic distance between nodes. The two balls in Figure 3.2(b) visualizes the probability measures m_{v_1} and m_{v_2}. All four nodes of the network in Figure 3.2(a) are the points on the "balls" m_{v_1} and m_{v_2}. The orange matrix in Figure 3.2(b) lists the probability measures of all nodes.



$d_{v_i v_j}$	v_1	v_2	v_3	v_4
v_1	0	1	1	2
v_2	1	0	1	2
v_3	1	1	0	1
v_4	2	2	1	0

$m_{v_i}(v_j)$	v_1	v_2	v_3	v_4
m_{v_1}	0	1/2	1/2	0
m_{v_2}	1/2	0	1/2	0
m_{v_3}	1/3	1/3	0	1/3
m_{v_4}	0	0	1	0

m_{v_1}	m_{v_2}		
	$\gamma^*(v_x, v_y)$	v_1	v_3	Sum
	v_2	1/2	0	1/2
	v_3	0	1/2	1/2
	Sum	1/2	1/2	1

Figure 3.2: (a) A simplified network with four nodes $V(G) = \{v_1, v_2, v_3, v_4\}$ and $E(G) = \{v_1v_2, v_2v_3, v_3v_4, v_3v_1\}$. Each element in the green table shows the value of $d(v_i, v_j)$. (b) Illustration of the probability measures for the simplified network. The two balls centered at v_1 and v_2 visualizes the probability measures m_{v_1} and m_{v_2}. The orange table lists the value of $m_{v_i}(v_j)$. (c) Illustration of the Wasserstein distance calculation. We take the calculation of $w(v_1, v_2)$ as an example. The blue table shows the optimal transportation plan between probility measures m_{v_1} and m_{v_2}. (d) Illustration of the network embedding. The example network (in (a)) is embedded in a Riemannian manifold with Wasserstein distance. The numbers attached to each edge v_iv_j denote the Wasserstein distance $w(v_i, v_j)$ and Olliver-ricci curvature $OR_{v_iv_j}$. In all these three tables, darker color represents higher value in this element.

The distance between centers of "balls" m_{v_i} and m_{v_j} is defined as the geodesic distance between nodes v_i and v_j, i.e., $d_{v_iv_j}$. We now define the transportation distance between two "balls". Since each "ball" is a probability measure, we calculate the transportation distance between m_{v_i} and m_{v_j} using the well-known

L^1 Wasserstein distance, which is defined as follows,

$$w(m_{v_i}, m_{v_j}) = \inf_{\gamma} \sum_{v_x \in N(v_i)} \sum_{v_y \in N(v_j)} d_{v_x v_y} \gamma(v_x, v_y), \tag{3.5}$$

where $N(v_i)$ and $N(v_j)$ are the support of m_{v_i} and m_{v_j}, i.e., all adjacent nodes of node v_i and v_j, respectively; γ represents the transportation plan, which is a matrix with elements $\gamma(v_x, v_y)$ denoting the mass of moving from v_x to v_y, the infimum is taken over all possible γ. Note that γ satisfies that $\sum_{v_y \in N(v_j)} \gamma(v_x, v_y) = m_{v_i}(v_x)$ and $\sum_{v_x \in N(v_i)} \gamma(v_x, v_y) = m_{v_j}(v_y)$. The particular transportation plan which results in the optimal cost is called optimal transportation plan γ^*. Figure 3.2 (c) shows an example of calculating the Wasserstein distance between m_{v_1} and m_{v_2}. The support $N(v_1)$ and $N(v_2)$ are v_2, v_3 and v_1, v_3, respectively. The optimal transportation plan between m_{v_1} and m_{v_2} is shown in the blue matrix in Figure 3.2 (c). The Wasserstein distance is thus calculated using the optimal transportation plan (blue matrix in Figure 3.2 (c)) and the geodesic distance (green matrix in Figure 3.2 (a)).

The Ollivier–Ricci curvature $\kappa(u, v)$ for node pair u and v is formally defined as

$$\kappa(u, v) = 1 - \frac{W(m_u^\alpha, m_v^\alpha)}{d(u, v)} \tag{3.6}$$

where W_1 is the Wasserstein distance defined in Equation (3.5). The $\kappa(u, v)$ is positive if the distance between the probability measures m_u^α and m_v^α is less than the distance between u and v. Compared with the Ricci curvature of the Riemannian manifold, the Ollivier–Ricci curvature discards the scaling factor $\varepsilon/2n$. Also, in the Riemannian manifold, the Ricci curvature was defined along a tangent vector. Here, in a network, the Ollivier–Ricci curvature is defined along a pair of nodes which are close enough. In this paper, we consider two nodes as close enough if their shortest distance is no greater than two. Thus, we only calculate curvature for nodes whose shortest distance is no greater than two. Furthermore, when two balls strongly overlap, the transportation distance between the two balls is small, the Ricci curvature then has a large positive value. In the network case, when the two nodes have strong overlap between

neighborhoods of u, v, the transportation distance is small, and the Ollivier–Ricci curvature thus has a large positive value (Jost & Liu, 2014).

3.3 Ollivier Ricci Curvature Gradient and Community Structure

The OR curvature of the graph reveals the properties of the underlying Riemannian manifold and provides insights into the community structure of graphs. Previous work has shown that the large (or small) curvature corresponds to the edge which is more (or less) connected than a grid (Gosztolai & Arnaudon, 2021; Ni et al., 2015; Ni et al., 2019; Samal et al., 2018). As shown in Figure 3.3, the left is a graph with five communities, and the right one is a graph with three communities generated by the stochastic block model. The warmer color indicates a smaller curvature, and the colder color indicates a larger curvature. Previous empirical results and the above toy example have related the OR curvature of graphs to the community structure of the graph. However, rigorous theoretical proofs that the OR curvature of the within-community edge is larger than the between-communities edge in a graph are not developed yet.

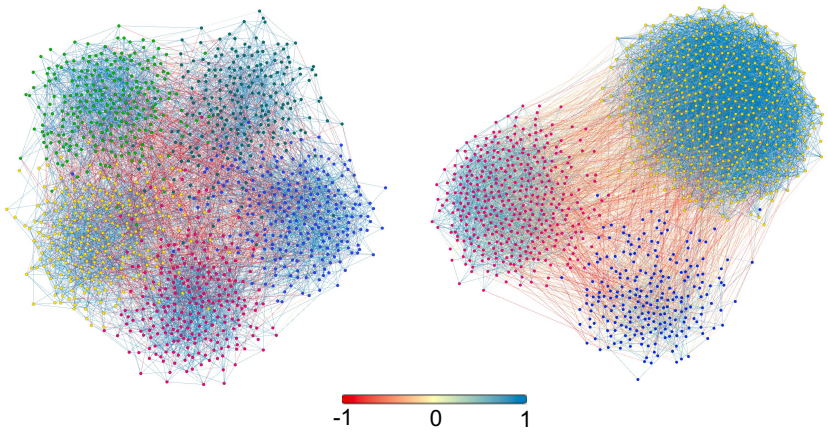


Figure 3.3: The cold-colored edges have large OR curvatures, and the warm-colored edges have small OR curvatures. Nodes with different colors belong to different communities.

3.3.1 OR Curvature Gradient and Communities

We define the OR curvature gradient (ORG) as $|\kappa(x^{(j)}, y^{(j)}) - \kappa(x^{(i)}, y^{(i)})|$, where $(x^{(j)}, y^{(j)}) \in \Delta((x^{(i)}, y^{(i)}))$. It is the difference between the OR curvature of two adjacent edges. According to the theoretical results, the within-community ORG $|\kappa(x_{B_i}, y_{B_i}) - \kappa(x'_{B_i}, y'_{B_i})|$ is significantly smaller than the between-communities ORG $|\kappa(x_{B_i}, y_{B_i}) - \kappa(x^*_{B_i}, y^*_{B_j})|$, this motivates us to propose a subsampling algorithm based on ORG, which extracts the community information with theoretical guarantee.

3.4 OR Curvature Gradient Based Graph Subsampling

The existing graph subsampling methods usually use degree information and random walk through the graph. However, the subsampled graph tends to leave out minority communities (i.e., a community with a smaller number) since they tend to sample nodes with high degrees. Due to the lack of taking advantage of the community structure information during the expansion, the available subsampling algorithms usually underestimate M. From the previous work, the community structure relates to the geometric phenomenon of the underlying Riemannian Manifold. Thus, we develop a fast and efficient ORG-based subsampling algorithm that is able to traverse the communities of the graph instead of trapping in a single community. We then prove the probability of the proposed ORG-sub taking all communities of the full graph into the subsample converges to one as the subsample size increases. Consequently, the subsampled graph attained by the ORG-sub algorithm can get a larger proportion of the minority communities than other subsampling algorithms and helps estimate M more accurately, as shown in Figure 3.4. In addition, we theoretically show the estimation of M by the subsampling algorithm converges to true M.

3.4.1 Subsampling Algorithm

The ORG carries the community structure information since it differentiates within-community edges and between-communities edges. We can use ORG as a guide for expanding the subgraph in the direction

of more communities. First, we randomly sample an edge $(x, y) \in E$ as the starting edge. The probability of obtaining an edge (x, y) at start is $P((x^{(1)}, y^{(1)})) = \frac{1}{|E|}$.

To take advantage of the theoretical properties of the ORG, the ORG-based subsampler expands to the next edge whose OR curvature shows the greatest difference from the OR curvature of the previously taken edge. The edge subsampled in the $i+1$-th step $(x^{(i+1)}, y^{(i+1)})$ given the edge subsampled in the i-th step $(x^{(i)}, y^{(i)})$ can be expressed by:

$$(x^{(i+1)}, y^{(i+1)}) = \underset{(x,y)\in\Delta((x^{(i)},y^{(i)}))}{\operatorname{argmax}} |\kappa(x, y) - \kappa(x^{(i)}, y^{(i)})|. \tag{3.7}$$

Since the difference between the within-community ORG is smaller than the between-community ORG, the proposed subsampler will expand to another community instead of trapping in one community. After the subsampler stops expanding or the subsampling budget is used up, we get subsampled nodes. Then we can get the subgraph induced by the subsampled nodes. Details of the algorithm are in Algorithm 2.

Algorithm 2

Input: Graph G; Number of nodes to be subsampled $\tilde{n}$; OR curvature of the graph G calculated by parameter α; the subsampled node set $S = \emptyset$
Initialization: Randomly choose a node v_0 as the start of sampling. Here, $t = 0$.
Cold Start: Among all neighbors of v_0, randomly add v_1 to S. Here, $t = 1$.
While: $|S| < \tilde{n}$

- Step 1: Given the edge $e_t = (v_{t-1}, v_t)$ selected in the t-th step, get the edge curvatures of the e_t's neighboring edges $\Delta(e_t)$.
- Step 2: Select the edge e_{t+1} in the edge set $\Delta(e_t)$ that has the greatest OGR.
- Step 3: Add nodes connecting the edge e_{t+1} to the subsampled node set S.
- $t = t + 1$.

Output: Subsampled node set S and the induced subgraph $G[S]$

3.4.2 ORG-sub for Estimating M

Applying the proposed ORG-sub algorithm to the full graph, as shown in Figure 3.4 (A), we can get a subsample as shown in Figure 3.4 (B). Compared with the subsample obtained by the degree-based subsampler in Figure 3.4 (C), the proposed algorithm can subsample more nodes in minority communities.

The toy example illustrates that the proposed ORG-sub algorithm preserves the community structure information better. Thus, the subsampled graph can help estimate M of the full graph.

We take multiple subgraphs by applying ORG-sub r times. For each subsampled graph $G[S_i]$ given the sample size $\tilde{n}$ at i-th time, we get an estimation of M, $\hat{M}(G[S_i])$, then the mean of $\hat{M}(G[S_i])$ is calculated as our final estimation of M. As for the choice of the estimation algorithm, we use the state-of-the-art network cross-validation method (Li et al., 2020).Compared with other methods (Chen & Lei, 2018; Y. R. Wang & Bickel, 2017), network cross-validation is more robust to complicated settings, i.e., the number of communities is large or the community structure is not conspicuous.

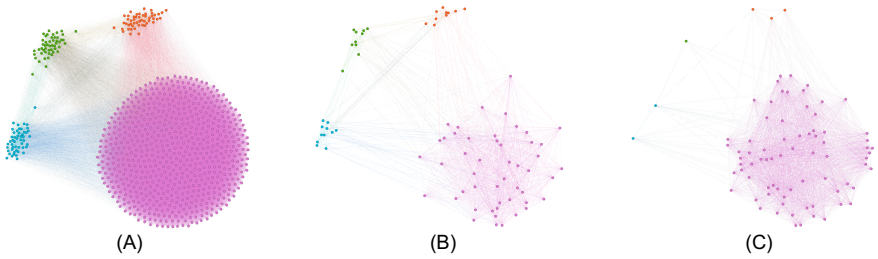


Figure 3.4: (A): A graph generated by SBM with community size $(650, 50, 50, 50)$, $p_{in} = 0.8$ and $p_{out} = 0.1$. (B): The subsampled graph by the proposed ORG-sub algorithm. The proportion of subsampling is 10%. (C) The subsampled graph by degree-based subsampling algorithm. The proportion of subsampling is 10%.

3.4.3 Computational Analysis of ORG-sub

Computational complexity of Algorithm 2. We analyze the time complexity of Algorithm 2 step by step. The time complexity of querying the neighboring edges and their corresponding curvatures is of order $\mathcal{O}(n\tilde{d})$ by Nys-sink algorithm (Altschuler et al., 2019), given $\tilde{d}$ is the average degree. The time complexity of the sorting step to attaining the most different curvature from the neighbors is $\mathcal{O}(\tilde{d}\log\tilde{d})$. Thus, the time complexity of taking one step is of order $\mathcal{O}(n\tilde{d})$. To sample $\tilde{n}$ nodes, we need to run about $\tilde{n}$ steps and get the induced subgraph from the sampled node set S. Since getting an induced subgraph of size $\tilde{n}$ takes time complexity of order $\mathcal{O}(\tilde{n}\tilde{d})$, the total complexity of the sampling procedure is $\mathcal{O}(n\tilde{n}\tilde{d})$.

Choice of the subsample Size. When implementing our algorithm, we need to specify the size of the subsamples, which determines the performance and efficiency of the algorithm. Empirically, we plot

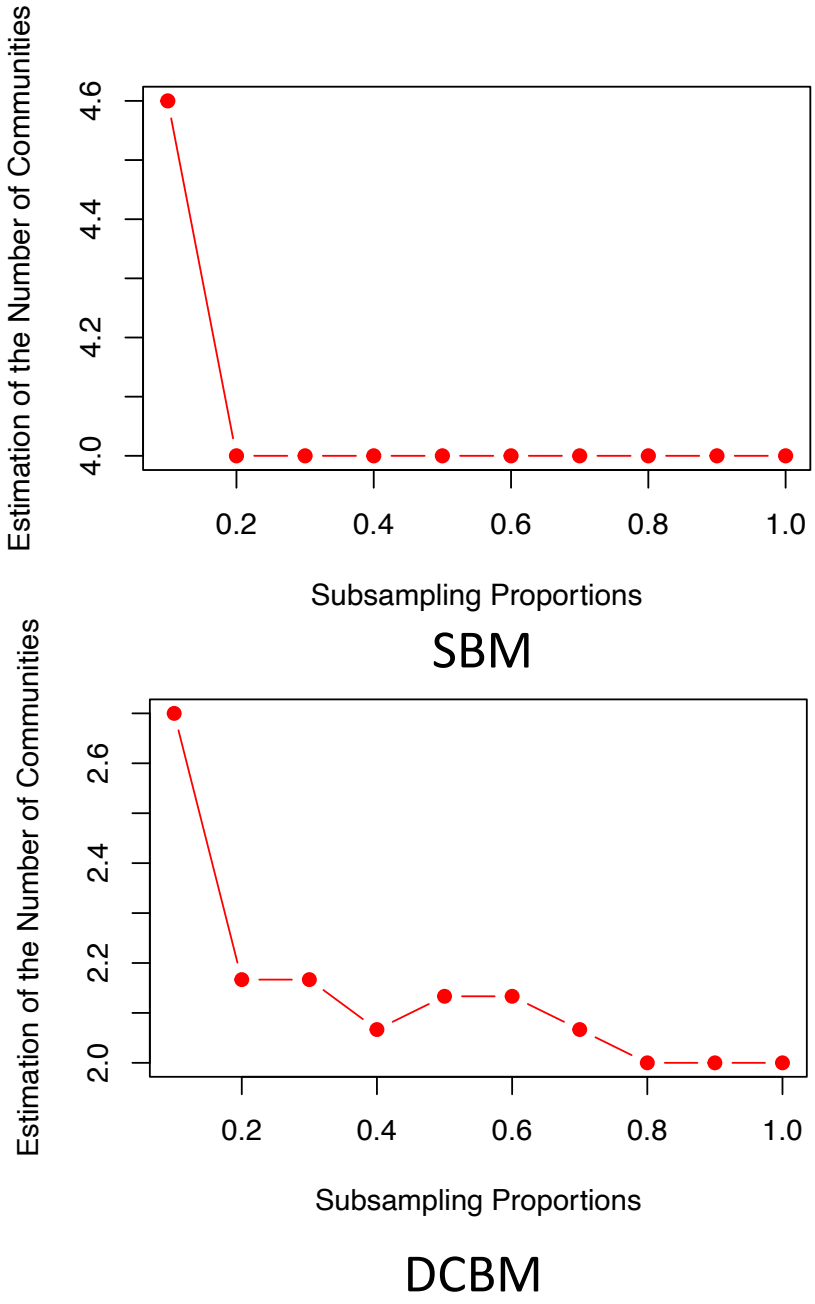


Figure 3.5: The function of the estimation of M with respect to the subsampling proportion for SBM and DCBM, respectively.

the estimation of M as the function of the subsampling proportion in Figure 3.5, and we can select the

elbow where the estimation of M starts to converge in applications to save computational resources and stabilize the estimation results.

3.5 Experiment

We evaluate the performance of our algorithm on both synthetic datasets and real-world datasets. We use the metric, MAE$= \frac{1}{r}\sum_{i=1}^{r}\left|M - \hat{M}(G[S_i])\right|$ to evaluate the accuracy of subsampling based estimation. We compare the proposed method with (1) Degree-based Node sampler (DBN) (Adamic et al., 2001), (2) Community Structure Expansion Sampler (CSE) (Maiya & Berger-Wolf, 2010) which is a community structure preserved sampler, and six benchmark exploration-based graph subsampling methods, including (3) Metropolis Hasting Random Walk Sampler (MHRW) (Hübler et al., 2008), (4) Forest Fire Sampler (FFS) (Leskovec et al., 2005), (5) Snowball Sampler (Goodman, 1961), (6) Random Walk Sampler (RW) (Gjoka et al., 2010), (7) Multi-dimensional Random Walk Sampler (MDRW) (Ribeiro & Towsley, 2010). As for other methods, we set the hyper-parameters as the default in package *Little Ball of Fur* (Rozemberczki et al., 2020b). We replicate the experiments 30 times under each setting and compare the performance in terms of MAE for all methods. All the experiments are conducted on a machine with a 40-core NVIDIA Tesla V100 GPU (3.00 GHz).

3.5.1 Datasets

Synthetic Dataset We generate synthetic datasets by the stochastic block models (SBM) and degree-corrected block models (DCBM), which can assign the community distribution to each node. We set the community proportion as $(3/4, 1/10, 1/12, 1/15)$ with 900 nodes in total. The out-in-ratio (the ratio of between-communities edges over within-community edges) controls the ratio of the number of edges between the communities and within a community. A higher ratio represents a noisier graph. The degree-corrected model corrects the node degree by a power-law distribution. Given the probability of an edge within a community ($p_{in} = 0.8$), we vary the probability of an edge between communities (p_{out})

($\{0.06, 0.08, 0.10, 0.12\}$) and subsampling proportions ($\{0.1, 0.12, 0.14, 0.16\}$) from low to high for both block models to observe how our method performs as the setting becomes more challenging compared with others. More details about the generation of the DCBM datasets are presented in Appendix B.

Real-world Dataset We use five widely-used read-world graph datasets with labeled community structures to validate the performance of our method, including Polbooks, Facebook, Cora, Polblogs, and PubMed (Rossi & Ahmed, 2015; Rozemberczki et al., 2020a; Sen et al., 2008). All these graphs are considered unweighted and undirected. Besides, all the self-loop edges and isolated nodes are removed. Appendix B summarizes the network statistics of these datasets. As we can see, the number of nodes ranges from 105 to 19,717, and the network density ranges from 0.0001 to 0.04. Thus the networks we considered span a wide range.

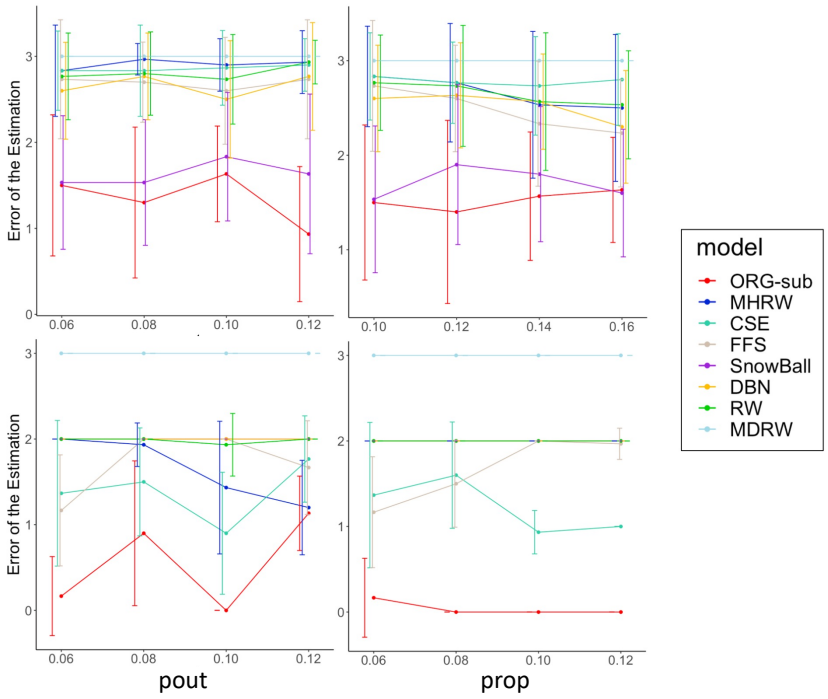


Figure 3.6: The function of the performance of MAE with respect to pout and prop for DCBM dataset.

3.5.2 Results of Synthetic Dataset

The subsampling times r for estimating M is set as 3, which is enough to get a stable estimation result (Li et al., 2020). The results of SBM and DCBM datasets are reported in Figure 3.6. The error bar is the standard deviation of 30 replications. We set the subsampling proportion (prop) as 0.1 when varying the probability of an edge between communities (pout) and set pout=0.06 when varying the prop. Our method can get a more accurate estimation of M than other methods. More details about the table of results, including each combination of pout and prop values, are presented in Appendix B. We also compare the computation time for estimating M by using the whole graph and the subgraph. It turns out the time for estimation of the subgraph together with the time for subsampling is still much shorter than the time for estimation on the whole graph. It is consistent with our conclusion about the complexity of the algorithm. Details about computational time are presented in Appendix B.

3.5.3 Results of Real-world Dataset

We compare the performance of the proposed ORG-sub algorithm with subsamples generated by different algorithms by MAE defined above. Table 3.1 records the average error of the estimation of M over 30 replications. Still, our algorithm outperforms other algorithms in most cases. Let $n_1, \ldots, n_M$ denote the number of nodes in M communities. Here, we calculate the normalized Shannon Entropy of $n_1, \ldots, n_M$ as a metric to measure the imbalance level, i.e., we calculate IM $= -\frac{1}{\log M} \sum_{i=1}^{M} \frac{n_i}{n} \log \frac{n_i}{n}$ and a higher IM value means that the communities are more imbalanced. Appendix B summarizes the IM. As we can see, PubMed and Polbooks are both more imbalanced than Facebook. From Table 3.1, we observe that the ORG-sub performances of PubMed and Polbooks are better than that of Facebook. This shows that ORG-sub performs better when the community of the graph is more imbalanced. The true M of each dataset is recorded in the first column of Table 3.1. We tested the performance of different algorithms with a broader range of sampling proportions ranging from 0.5% to 30%. In particular, for small-sized networks Polbooks and Facebook, we consider sampling proportions 10%, 20% and 30%. For medium-size networks

Polblogs and Cora, we consider 5%, 10%, and 20%. For large-size network PubMed, we consider 0.5%, 2%, and 5%. In addition, the computation time on the subgraph (together with the subsampling time) is still much shorter than on the whole graph. Details about computational time are presented in Appendix B.

Table 3.1: Comparison of the performance on the error of the estimation of M for each dataset and subsampling method.

Dataset	Prop.	ORG-sub	MHRW	CSE	FFS	SnowBall	DBN	RW	MDRW
Polbooks True: 3	10%	**0.00**	1.20	0.62	2.68	0.48	0.60	0.33	0.00
	20%	**0.00**	0.19	0.52	0.30	0.70	1.76	0.37	1.60
	30%	**0.23**	1.00	0.43	0.37	0.93	1.60	0.37	2.00
Facebook True: 11	10%	**5.27**	6.83	7.97	6.87	8.77	7.57	5.77	10.00
	20%	5.67	5.13	5.93	6.93	7.20	**4.23**	5.50	9.90
	30%	**5.90**	6.20	8.97	7.13	8.77	7.27	8.97	10.00
Cora True: 7	5%	2.27	2.33	4.93	5.97	3.77	3.10	**1.30**	4.60
	10%	3.40	**1.53**	4.97	3.40	3.00	3.37	3.87	5.80
	20%	**2.53**	2.97	5.00	3.90	4.80	4.80	2.73	5.47
Polblogs True: 2	5%	**0.00**	1.87	0.90	2.00	0.43	1.33	1.03	0.30
	10%	**0.00**	0.40	0.33	0.20	0.03	0.03	0.07	0.87
	20%	**0.00**	1.87	0.90	2.00	0.43	1.33	1.03	0.30
PubMed True: 3	0.5%	**0.20**	0.30	0.30	0.50	1.00	0.70	0.30	1.90
	2%	**0.00**	0.30	0.80	0.40	0.20	0.70	1.20	1.80
	5%	**0.30**	0.55	0.55	1.40	1.65	0.65	2.20	2.75

3.6 Conclusion

In this paper, we propose a novel Ollivier-Ricci Curvature Gradient-based graph subsampling (ORG-sub) method that samples a subgraph by maximizing OR curvature gradient. The contribution of the ORG-sub method to the graph subsampling research line is three-fold. First, to the best of our knowledge, we are the first to utilize the graph's internal topological information to subsample a large graph that preserves the number of communities. Second, to the best of our knowledge, we are the first to bridge the gap in the consistency theory regarding subsampling algorithms for the number of community estimations in SBMs. In particular, we theoretically show that ORG-sub effectively traverse different communities and avoid trapping in one community. In addition, we theoretically show the advantage of ORG-sub over

a popular sampling method, random walk. Third, we empirically show that our method has superior performance over existing subsampling algorithms in terms of estimating the number of communities.

An interesting future direction we plan to investigate is to extend the theory of our method to SBM variants, including degree-correlated SBM, overlapping SBM, and multi-layer SBM. Another future direction is to investigate other community-related statistics that our method has preserved. In fact, we have empirically observed some promising results in preserving the clustering coefficient (CC). Empirical results can be found in Appendix B. In the future, we will investigate the consistency theory of CC.

Chapter 4

Analysis of Transnational Advocacy Network in Political Science

4.1 Introduction

Transnational advocacy networks (TANs) are the "most familiar example" of networks in international relations (Hafner-Burton et al., 2009). Made up primarily of non-governmental organizations (NGOs) from around the world, the information, resources, and power of transnational advocacy actors have been argued to increase as a result of their networking behavior (Risse-Kappen et al., 1999). By joining forces around a common cause, NGOs succeed in their collective struggle against repressive states or polluting governments, among the myriad of other advocacy causes.

While much research has shown how networking improves advocacy outcomes, other research has highlighted how the structure of advocacy networks can reinforce global power discrepancies. Even though NGOs are often assumed to increase governance access to the world's powerless, not all NGOs have equal access to or equal involvement with the advocacy network. Organizations from the global South are

regularly left out of the overall advocacy network (Fowler, 2000; Shumate & Dewitt, 2008; Townsend & Townsend, 2004). A lack of resources, both human and material, often means that organizations from the global South cannot attend NGO conferences or participate in working groups. When organizations from the global South do participate in the network, many raise concerns about exploitation by their global North counterparts (Kassal et al., 2008; Nelson, 1997). Organizations in the global North may use information on the plight of individuals in the global South for their personal fundraising or mission goals (Bob, 2005; Pallas & Urpelainen, 2013). Moreover, power differentials within the advocacy network may make it difficult for organizations in the global South to have their issues championed by the overall network (Carpenter, 2014).

The troublesome power disparities identified in the structure of the advocacy network are incredibly important in today's political environment. NGOs are currently facing a global backlash and "closing space" in the world community (Brechenmacher, 2017). Repressive regimes have ousted international NGOs for not being well connected to local communities and pursuing their own agenda. Regimes have stopped international aid to NGOs, claiming that this aid makes global South organizations beholden to the foreign policy desires of powerful countries from the global North. The power disparities that seem endemic to the advocacy network could diminish the access NGOs have to certain affected populations and/or provide a cover through which states could counter the advocacy concerns of civil society actors (Terman, 2019).

Our study seeks to contribute to this growing debate about the utility and structure of the advocacy network for NGOs:

- How power disparities influence the structure of the advocacy network
- How the community detection helps us understand emergent divisions within the advocacy network

- How concepts of brokerage and brokerage roles analysis help us understand how the transnational advocacy network can both extend power to organizations, while at the same time reinforcing power inequalities among NGOs

However, the following chapter is organized as follows. First, we outline the existing literature on transnational advocacy networks, paying special attention to divisions within this literature. Next, we present a cross-disciplinary review of the concepts of community and brokerage and apply these concepts to the study of NGOs. We present our argument and some testable hypotheses that flow from our logic. We then present our novel data and describe our results. Our study concludes with some practical steps that the UN and other concerned actors can take to increase representation and limit divisions within the NGO network.

4.2 Promise and Problems of TANs

When compared to the states and corporations that they often seek to change, NGOs are relatively powerless actors in international politics. They do not have standing militaries or deep coffers; many have no paid staff at all. NGOs gather information, frame issues, educate populations, and mobilize global dissent in order to pressure more powerful actors to change behaviors and adopt certain policies, like getting a government to release a political dissident or getting a company to reduce its carbon footprint. As the number of NGOs began to exponentially grow at the end of the Cold War, scholars recognized that organizations do not work in isolation (Brysk, 1993a). Successful advocacy often depends on organizations working together with a variety of other actors that share their same advocacy goals. These actors may include local movement leaders, churches, labor unions, parts of various intergovernmental organizations, and sympathetic government officials. Either internationally or domestically, organizations can also join forces with other NGOs in order to amplify their message and increase their reach.

In their study of international advocacy, Keck and Sikkink (M. E. Keck & Sikkink, 1998a) call this "dense web of connections" between organizations the transnational advocacy network. In certain situ-

ations, the transnational advocacy network increases the advocate's power, leading targeted actors to make concessions in line with the advocate's desires, even if those targeted had previously resisted change. One way that this network can work is through a "boomerang pattern," where connected domestic advocates call out to their transnational network partners to increase international pressure and domestic resources. As DeMars (DeMars, 2005b) points out, the overall advocacy network provides additional resources to involved organizations, giving them new tools through which to advocate for change. The transnational advocacy network, with NGOs as the key actor, can be thought of as a public good (Shumate & Dewitt, 2008). It increases the advocacy output and perceived success of involved organizations (A. Murdie, 2014; Tallberg, Dellmuth et al., 2018)

Keck and Sikkink (M. E. Keck & Sikkink, 1998a)'s characterization of the transnational advocacy network marked a turning point in the study of international politics. Unlike traditional realists, Keck and Sikkink (1998) and later work by Risse, Ropp, and Sikkink (1999), among others, laid out how NGOs and other advocates can become powerful players on the world stage, even without substantial military or material resources. Through this work, even the term "networks" became associated with a structure through which "otherwise weak actors" could "voice their interests and influence governance outcomes in international relations (Avant & Westerwinter, 2016). To note, most of the first wave of work that followed Keck and Sikkink (M. E. Keck & Sikkink, 1998a) took a "network-as-actor" approach, where "the network is no longer just a way of describing relationships among actors, but an actor until itself" (Kahler, 2011).

The study of the transnational advocacy networks then moved from a "network-as-actor" and to a "network-as-structure" approach, examining relationships within the network (Kahler, 2011; A. Murdie & Polizzi, 2017). This shift brought many critiques against the somewhat rosy view of NGOs and transnational advocacy networks that had dominated the turn of the millennium. Some of these critiques focused on the assumption that NGOs were more "principled" than other actors in world politics (Cooley & Ron, 2002). Other, more "network-as-structure" critiques centered on how advocacy networks might mirror global power inequalities, with powerful organizations from the global North controlling the advocacy

network for their personalistic goals. NGOs in the global South can become dependent on Northern organizations with more preexisting resources, or may be missing from the overall network in the first place (Jackson, 2020; A. Murdie, 2014). Savvy organizations and affected populations may have to market their plight to global North organizations interested in their own personal gain (Bob, 2005). Moreover, powerful organizations can act as "gatekeepers," keeping new ideas and issues from permeating through the network (Carpenter, 2014). In addition, the network can include organizations that are free-riders, taking information and resources from others without contributing to the public good (A. Murdie, 2014). The structure of the network can also affect the strategies taken by organizations, sometimes limiting innovation (Bush, 2015; Hadden & Jasny, 2019; Wong, 2012). To the best of our knowledge, there have been few attempts to reconcile the optimistic view of many of the "network-as-actor" studies with the more pessimistic "network-as-structure" research.

We argue that the network science concepts of community and brokerage provide us with a rich theoretical lens through which to understand the complex nature of the transnational advocacy network. These concepts are relatively new to the growing network literature in IR, especially the literature on NGOs. Below, we first outline how community detection helps us understand divisions within the advocacy network. We then turn our focus to brokerage, a concept which can greatly help us understand how the transnational advocacy network can both extend power to organizations, while at the same time reinforcing power inequalities among NGOs.

4.3 Communities in Network Science

Although sometimes implicit, TAN scholarship has not envisioned one advocacy network but many separate network structures. Networking is costly; it involves sharing information or resources for the common good. As such, an organization should only connect to others when it can benefit from the network partners' ideational and material resources. Over time, networking costs and benefits should lead to not one "dense web of connection" for NGOs but multiple distinct subnetworks unified over shared values, ideas, or targets (M. E. Keck & Sikkink, 1998a).

This basic understanding of the creation of transnational advocacy networks molds well with ideas from network science, especially with research on community detection (Cheng et al., 2018; M. E. Newman, 2006). Networks have emergent properties: as connections occur, new subgroups can emerge within the structure (Maoz, 2017). These endogenous groups or communities are determined both by the actors themselves but also by the evolving structure of the network.

A community is defined as a "group of nodes that are more tightly connected to each other than they are to the rest of the network" (Liu et al., 2019; Mucha et al., 2010). For our purposes, nodes refer to the organizations in the network. Moreover, for our purposes, the network can be thought as the web of connections between organizations generally. Within this overall network, we can identify certain communities that are closely connected. These communities may be driven in part by shared characteristics or issue; however, the communities are also endogenously driven. As Maoz (Maoz, 2017) remarks, these emergent communities form "naturally" over time as a result of the ties that actors develop.

In our empirical models, we rely on a community detection algorithm that will endogenously determine the best composition of communities in our evolving advocacy network over the years (You et al., 2016). Community detection, which is increasingly common in network science, is succinctly summarized by Newman (M. E. Newman, 2006):

"Community structure methods normally assume that the network of interest divides naturally into subgroups and the experimenter's job is to find those groups. The number and size of the groups are thus determined by the network itself and not by the experimenter"

Community detection insights do not provide us with testable hypotheses per se, other than the one crucial implication that there will be distinct communities identified in the overall network (Implication 1). For us, community detection is similar to data coding. We use community detection methods to identify endogenous communities in the network. Descriptive and qualitative research then help us interpret the validity of the assignment of NGOs into the communities identified.

Unlike most existing work on NGO networks, we do not assume that communities are driven only by broad issuefocus. Community detection methods allow us to remain open to endogenous complexity in

community formation. Communities may be driven in part by shared characteristics or issue focus, but they also may form for a myriad of reasons particular to the evolving network structures. Organizations observe the network structure and the makeup of emerging communities and self-select into ties that reflect this structure and their interests. For example, NGOs may examine the network structure and make connections based on where they think they could be the most useful or where they think they could benefit the most from the existing pattern of ties. NGOs may want to join communities that help them address the problems and challenges they are concerned about, even if these problems straddle issue divides. In this way, we think the community detection approach captures the theoretical insights of the classic "network-as-actor" TAN scholarship and fits with existing conceptualizations of communities in IR, especially work on transnational communities of practice.

4.4 Communities in International Relations

Communities have been studied in many distinct ways within IR (E. Adler, 2005; Deloffre, 2016; M.-L. Djelic & Quack, 2010b; Hutchison, 2016; Schneiker & Joachim, 2018; Tsingou, 2015). Perhaps one of the most well-known examinations of communities in IR is the study of epistemic communities. Epistemic communities comprise knowledge-based experts that come together for problem-solving and advocacy. Epistemic communities, security communities, and TANs are "different interpretations" of a broader term used in many social sciences, that of "communities of practice" (E. Adler, 2005).

Communities of practice are groups of people who share a concern or passion and learn how to do it better as they interact regularly" (Wenger, 2006). The term originated in studies of apprenticeship and learning (Lave & Wenger, 1991). Communities of practice discover and share resources within a group that is focused on solving common problems (Koliba & Gajda, 2009). Adler (P. Adler, 2005) asserts that TANs are really communities of practice because a knowledge domain—for example, human rights—constitutes their likemindedness and practices." The conceptualization of a TAN by Keck and Sikkink (M. Keck & Sikkink, 1998) also fits within the literature on transnational communities, a communitarian concept that includes communities of practice (M. Djelic & Quack, 2010, 2011). These concepts have already enriched

studies of TANs and NGOs. For example, communities of practice insights have helped to explain social learning among NGOs in South East Asia and the deepening of global accountability communities of NGOs over time.

A greater focus on transnational communities of practice could further our understanding of NGOs and TANs in many ways. First, communities of practice scholarship allow us to connect to the practice turn "within IR, focusing on the process of collective knowledge in action (Bueger & Gadinger, 2015). More attention to knowledge creation and dissemination practices may provide insights into how certain NGO practices, like shaming, became popular among some NGO communities and stayed popular even in locations where there was evidence of shaming backlash (Snyder, 2020). In line with the communities of practice literature, shaming is a practice that may continue even when it starts to become unhelpful; it could take time to displace oldtimers" within a community of practice to allow for innovation (Lave & Wenger, 1991).

Second, communities of practice do not necessarily have fixed memberships. When transnational in nature, their members can be highly dispersed …across a multiplicity of countries that are rarely in direct contact" (M. Djelic & Quack, 2010). The fluid" nature of transnational communities can be a source of strength and innovation (M.-L. Djelic & Quack, 2010a). Fluidity implies that actors can be involved with others in multiple communities (P. Adler, 2005). In existing studies of TANs and NGOs, we typically do not allow for fluid movement. Instead, potential involvement in the network is assumed based on the issue area. Further, studies on transnational communities of practice have focused on how "node-to-node connections get progressively embedded and set" over time (M.-L. Djelic & Quack, 2010a). TAN scholarship has not historically focused on change over time, eliminating our ability to theorize and study how connections and network structures evolve within a community of practice to allow for innovation (Lave & Wenger, 1991).

We contend that community detection from network science is a powerful and transparent way to begin identifying the boundaries of communities of practice. Community detection algorithms assume that communities are endogenous and evolving. Community detection methods help us find the bound-

aries between distinct communities, even given this fluidity. Further, by joining ideas of community with views of brokerage and brokerage roles, we are better able to understand the potential overlap between communities. Even beyond the study of NGOs and TANs, community detection and brokerage insights have much potential to help enrich studies of other IR topics, including communities of practice.

4.5 Brokerage in networks

If TANs comprise endogenously emerging communities of practice, how can we understand the power disparities identified in the existing TAN "network-as-structure" literature? How can the network be a powerful actor for international political change, as identified in the TAN "network-asactor" literature, and reinforce or exacerbate existing power inequities, as outlined in the later "network-as-structure" approach? We argue that network science and sociological discussions of distinct brokerage roles are vital to understanding this puzzle.

Brokerage is defined as a relationship "involving three actors, two of whom are the actual parties to the transaction and one of whom is the intermediary or broker" (Gould & Fernandez, 1989b). We have brokers in many different aspects of our daily lives, from the real estate agent that brokers a deal between buyer and seller to the department head that serves as an intermediary between fussy colleagues. Brokerage is thought to help with innovation (Avant & Westerwinter, 2016). When focusing on advocacy networks, we can think of the broker as the organization that connects two NGOs that otherwise would not have been connected. This could be the regional NGO that takes the interests of a local NGO and frames it for presentation to a big international NGO. It could be the sustainable development organization that participates in discussions both with health and with environmental organizations. Alternatively, it could be the organization that reaches out individually to transitional justice advocates and women's rights advocates after it hears of the abuse of a local individual. This basic idea is at the heart of our understanding of transnational advocacy networks, specifically the boomerang model (M. E. Keck & Sikkink, 1998a). There are organizations that help in connecting those with needs to those with resources, both internationally and domestically.

Sociologists have highlighted a "dual aspect of brokerage" (Stovel & Shaw, 2012). While brokerage does help transmit information and could help in the pursuit of a specific social, economic, or political goals, brokerage "often breeds exploitation, the pursuit of personal profit, corruption, and the accumulation of power," exacerbating "existing inequalities" (Stovel & Shaw, 2012). Brokers can accumulate power as they control information and access between competitive groups (Burt, 2003). Brokers may control access to communities in ways that stifle innovation. They could selectively relay information for their own goals, ultimately accumulating more power from their brokerage roles. As Stovel and Shaw (Stovel & Shaw, 2012) point out, the idea that brokers benefit from their role is well-established in sociology and network science; brokers can gain "money, information, access to opportunities, enhanced status, or ill-defined claims on side parties' loyalty". This discussion of brokers' accumulation of power is very similar to the "network-as-structure" critiques about transnational advocacy networks: certain organizations enhance their power from the network, acting as gatekeepers or using network information for their gain (Bob, 2005; Carpenter, 2014; Jordan & Van Tuijl, 2000).

On the other hand, new research from organizational studies has shown that not all brokers accumulate equal benefits from their brokerage role; certain brokers may lose status in the eyes of their peers. Because brokers are conduits between other actors, they may receive little attention and lack a clear identity themselves, ultimately hurting their status (Sullivan & Stewart, 2017). Crucially, however, the possible negative effect of the brokerage on actor status may depend on the "actor's prior established status" (Sullivan & Stewart, 2017). Actors without a prior high status may see brokerage reduce their status, unlike the general idea that brokerage is a conduit of social power seen in much of the sociological research.

Gould and Fernandez (Gould & Fernandez, 1989b)'s influential work serves to connect ideas of community and brokerage. There are different roles that brokers take dependent on whether they are acting as a broker within a specific community or between communities. According to their influential work, in a nondirectional network, there are four potential brokerage roles: coordinator, itinerant broker, gatekeeper/representative, and liaison. Figure 4.1 provides a graphical representation of these relationships.

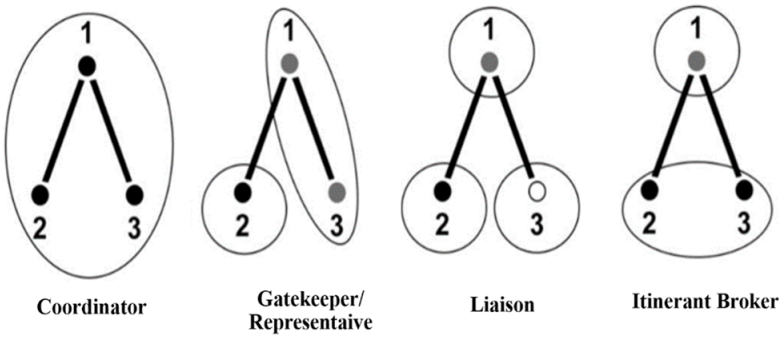


Figure 4.1: Brokerage roles

A coordinator, also called a local broker, is a broker within one specific community (Gould & Fernandez, 1989b). Within its community, it may have more resources or incentives to network than its peers. For NGOs, we could think of this as the organization that knows everyone within a specific area. Its leadership could have a long history working for different NGOs within an area (Henriksen & Seabrooke, 2016). The NGO could transmit ideas or share strategies between groups that might have trouble connecting directly. The status benefits that this coordinator receives for its role may only be known within the community.

An itinerant broker, also called a cosmopolitan broker, is a broker that is outside a community but is connecting members within that community. An itinerant broker may have special qualities or resources that enable it to connect to two actors within the same community, even when those actors cannot connect themselves. This could be the international NGO that serves to connect two domestic organizations at different sides of a country. This could be the NGO funder that has connections with two competing organizations that rely on outside donations (Cooley & Ron, 2002). Or, this could be an organization that works to represent itself to others outside its area of expertise, like a professional organization (Boli & Thomas, 1999).

A gatekeeper/representative broker connects its community to the outside community. It could be a gatekeeper, being the conduit of information into the community. Or, it could be the representative, being the conduit of information outside the community. Because it controls information and resource flow for members within its community, a gatekeeper/representative broker can accumulate a lot of power

and status. It may also be able to manipulate information and resource flows in ways that are personally beneficial. It may develop special skills that help it retain its role. Although not always using these terms, the gatekeeper/representative brokerage role has been talked about extensively in the "network-as-structure" literature on transnational advocacy. This is the NGO that can either facilitate or stop a new issue from making it to the broad advocacy stage (Bob, 2005; Carpenter, 2014).

Finally, a liaison broker connects two actors from separate communities, itself not being a member of either community. Mediators between unions and management would be a classic example of a liaison broker (Fernandez & Gould, 1994). There may be structural reasons why actors from each community cannot or do not connect directly. The liaison broker may have interests that bridge the communities, or it may have resources that allow it to connect when others do not. Liaison brokers may be capable of talking to diverse audiences about a moderate message, like Stroup and Wong (Stroup & Wong, 2017)'s discussion of "leading" international NGOs. According to Stroup and Wong (Stroup & Wong, 2017), leading organizations "receive deference from difference audiences in global politics and therefore have authority". These audiences "can be quite diverse in their preferences," requiring the leading NGO to develop a moderated approach to advocacy (Stroup & Wong, 2017).

We think unpacking brokerage into these distinct brokerage roles can be incredibly useful for international relations. To our knowledge, Gould and Fernandez (Gould & Fernandez, 1989b)'s conceptualization of these roles has received limited attention within our subfield.

4.6 Community and Brokerage: Implications for NGO and TAN Research

By combining network science and social science understandings of emergent communities, transnational communities of practice, and brokerage roles, we are able to build a richer picture of TANs and NGO-to-NGO networking with many novel empirical implications.

Our theoretical logic is straightforward. First, we assume that networking is costly. NGOs with more resources should be more likely to be part of the advocacy network in the first place. As previous work has shown, organizations with fewer resources, like often those from the global South, may be left out of the advocacy network (Hadden & Bush, 2020; A. M. Murdie, 2014b; Shumate & Dewitt, 2008). Although many initiatives have tried to make it easier for global South NGOs to network, power disparities persist over time (Okumu, 2019).

Second, we assume that network ties can be beneficial for an NGO and the causes it represents. This assumption is consistent with the "network-as-actor" TAN literature. Networking provides tools for NGOs to use in their work and helps their efforts for social and political change (Brysk, 1993b; DeMars, 2005a; M. E. Keck & Sikkink, 1998b). This assumption is also consistent with recent empirical literature that has found that networking leads NGOs to produce more output and successfully advocate with intergovernmental organizations (A. M. Murdie, 2014a; Tallberg, Sommerer et al., 2018).

Because networking is costly but potentially beneficial, we contend that NGOs should network strategically, only expending resources for networking when they expect the benefits to outweigh the costs. Over time, strategic networking behavior will lead to the development of not one overall advocacy network but distinct communities. As NGOs work collectively to solve problems that they are passionate about, these communities will develop their own shared practices and knowledge (E. Adler, 2005). Communities of NGOs will emerge endogenously, creating separation within the advocacy space. Community detection methods from network science will help us examine this simple but powerful implication: There will be distinct communities that emerge in the overall NGO network (Implication #1).

To note, community detection is not a method directly for hypothesis testing. However, building both on the literature on community detection and communities of practice, our theoretical logic implies that we should see clear subgroups in the overall network. If TANs did not behave as communities of practice and networking was equally beneficial for NGOs regardless of the evolving structure of the network, we would not expect distinct communities to emerge; there would be just one community of all NGOs in the network. Instead, we contend that the nature of NGO networking and the need for specialized

knowledge and practices will lead to multiple emergent communities. In line with this implication, the first goal of our empirical analysis is to identify emergent communities in the overall NGO network. Through investigating the detected communities qualitatively and descriptively, we can better understand the endogenous process underlying divisions in the NGO world.

As communities evolve, we should see distinct practices develop. Communities will adopt strategies that work best for their specific problems and dynamics. At the same time, NGOs with specific skills and habits will be drawn to certain communities. This implies that networking itself is a practice of the endogenously evolving community. Each community will develop distinct practices concerning how NGOs in their community connect with other NGOs inside and outside their community.

The typology of brokerage roles of (Gould & Fernandez, 1989a) allows us to further examine how community structure can influence the nature of network ties. Some communities may have problems and practices that lead to more NGOs in coordinator brokerage roles, rarely venturing out of the specific community. Other communities may have practices and goals that facilitate connections across communities in distinct ways, leading to more itinerants, gatekeepers/representatives, or liaison brokers. Over time, NGOs in specific communities will develop similar practices, leading different communities to value certain brokerage roles more and attract organizations with the social power to retain these specific brokerage roles. Community-level differences in brokerage role patterns will both reflect and contribute to growing power disparities within the overall network. This logic implies that there will be an association between communities and brokerage role distributions: The emergent communities will have different and distinct distributions of brokerage roles (Implication 2).

As mentioned, brokerage roles can both provide power and favor the powerful. Although existing work has found that the TAN may provide resources for all NGOs involved, these resources are not necessarily distributed evenly. Some organizations may potentially gain more power as a result of their particular brokerage roles. Similarly, it may take certain resources to gain specific brokerage roles, especially those roles that connect actors across different communities. We argue that preexisting power disparities between NGOs in the global North and the global South provide an opportunity for some organizations

to take on certain brokerage roles more easily. In particular, we contend that becoming an itinerant, a gatekeeper/representative, or a liaison broker requires both human and material resources that can be extended beyond one community. Moreover, organizations in intercommunity brokerage roles may then try to limit the ability of other organizations to take their social power, something found in previous work on gatekeeping in the NGO network. Conversely, remember that a coordinator broker connects actors within one community. Although this brokerage role would still favor the more powerful, becoming a coordinator broker may take comparatively less resources than other brokerage roles. As such, organizations without preexisting power may be more likely to take on this intracommunity brokerage role than the various intercommunity brokerage roles identified in (Gould & Fernandez, 1989a). This logic suggests that global North organizations may not only dominate the NGO network, as previously found, but that they may also dominate certain brokerage roles within and between emerging network communities. This implies: There is an association between being based in the global North and the specific brokerage roles adopted by NGOs in emergent communities (Implication 3).

Although straightforward, these implications help explain how networking by NGOs could improve advocacy through the creation of distinct communities while simultaneously exacerbating power inequalities through the distribution of brokerage roles and the power that can be obtained in some of these roles. Distinct communities may help NGOs develop repertoires of knowledge and tactics that help achieve shared goals. NGOs that connect these communities through specific brokerage roles can utilize their brokerage roles and their preexisting power to further inequality within the NGO network. As such, even when efforts are made to lower the barriers of access for organizations from the global South, the nature of the NGO network and its structure can still create a hierarchy. These ideas are thus opposed to traditional notions of the network as an alternative to hierarchical modes of organization.

4.7 Data Collection

Although theoretically important, NGO networking data has been particularly difficult to gather. Scholars have used many innovative sources and approaches (Bush, 2015; Carpenter, 2014), and data from the Yearbook of International Organizations (Caniglia, 2001; A. Murdie, 2014).

We take a somewhat different approach that allows us to examine the NGO network over time. Specifically, we crawled information provided on the UN's "integrated Civil Society Organizations (iCSO) System" during the summer of 2018 for NGO profiles.This is a database of over $24,000$ NGOs from all UN member states. Organizations opt into the database when they establish a relationship with the UN's Department of Economic and Social Affairs (DESA) or when they apply for consultative status with the Economic and Social Council (ECOSOC). Participation in UN meetings as an NGO can also lead to an organization having a listing in the database. As DESA's NGO Branch summarizes on its website, in addition to listing an overall organizational objective, NGO profiles include "a general part (name, address, organization type), contacts and meeting participation, activities, and information related to the substantive areas of DESA" (DESA 2019). Figure 2 provides a screenshot of a sample organization's profile. Figure 4.2 provides a screenshot of a sample organization's profile.

Figure 4.2: Screenshot of iCSO organizational profile

To obtain NGO-to-NGO network data, we first focused on the meetings that NGOs have attended over time. Figure 4.3 provides a screenshot of meeting participation for a sample organization.

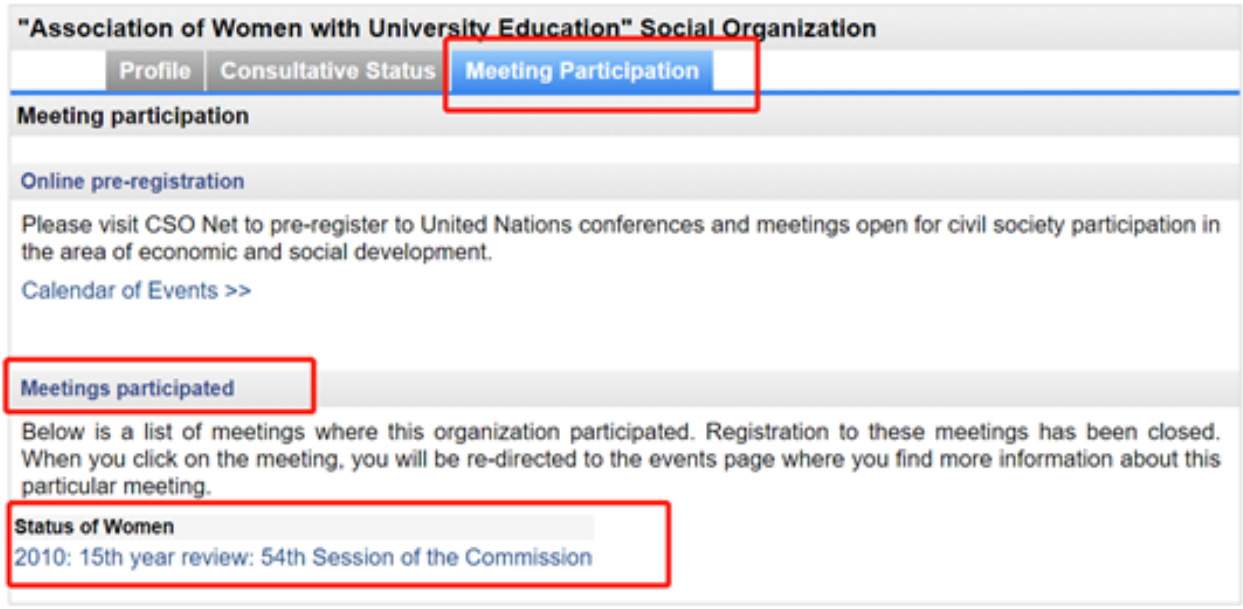


Figure 4.3: Screenshot of iCSO meeting participation

During our coding, we were able to identify meetings NGOs attended from 1992 through 2017. We then turned this two-mode (organizations and meetings) network dataset into a one-mode network of organizations connected through joint meeting attendance (Wasserman, Faust et al., 1994). This is common in community detection (Alzahrani & Horadam, 2016). In total, there were $3,903$ organizations and $1,300,519$ ties or edges. For our analysis, we deleted edges with a weight of one6 and then isolated nodes were removed, giving us a working dataset of $1,200$ organizations, with edge total varying by year. There were $437,152$ edges identified in the final year of the sample. Figure 4.4 shows how the number of connections between our nodes increases over time. After removing isolated nodes and edges with a weight of one, the remaining NGO organizations rarely registered on the UN's "integrated Civil Society Organizations (iCSO) System" before 2002. As a result, the following analysis focuses on the data from 2002 to 2017.

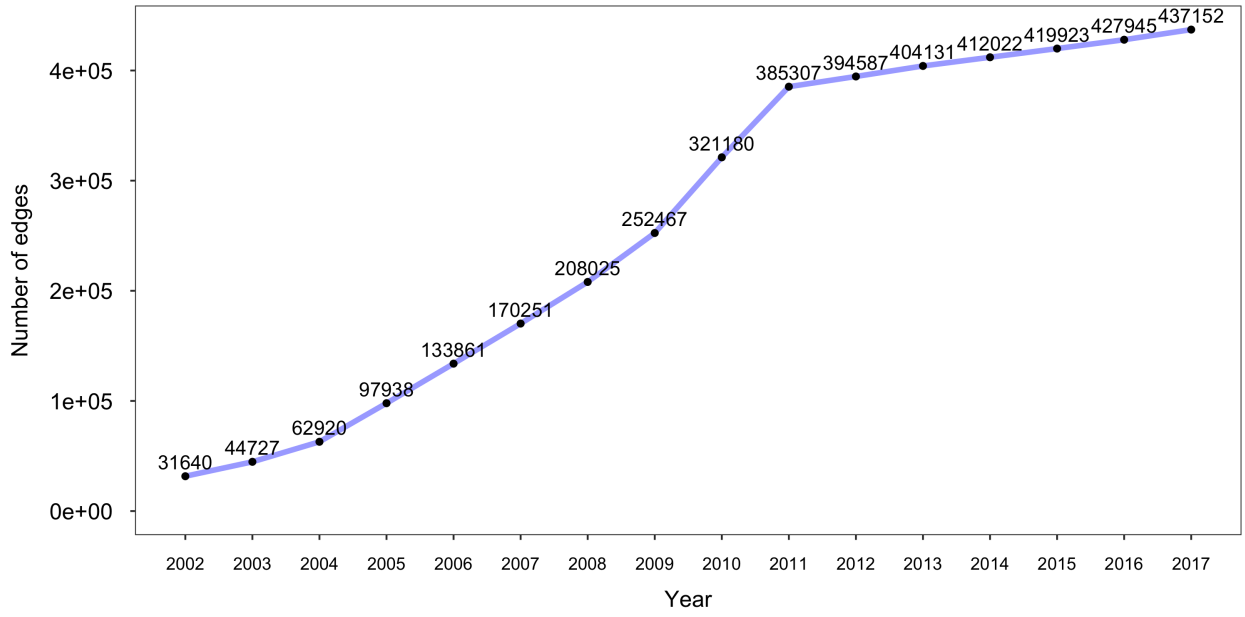


Figure 4.4: The NGO network over time

We think this dataset will be incredibly useful for future researchers. We also think it matches closely with our central concepts of interest. Joint meeting attendance provides an opportunity for NGOs to strategize, share information, change tactics, address targets, and form partnerships, all practices com-

monly thought of as NGO networking. Additionally, we think this dataset provides an especially stringent test of our hypotheses. As Moloo (Moloo, 2011) points out, the UN has taken steps over the time period of our sample to increase and ease access for NGOs in the UN system, especially relevant for organizations from the global South. The UN's legitimacy in global governance rests on NGO involvement (Moloo, 2011). As such, the costs of networking for global South organizations in UN meetings may be lower than other comparable networking forums, increasing their likely involvement and biasing our dataset against finding evidence for Hypothesis 1 and 3. Despite this, as discussed below, we find support for our three hypotheses.

4.8 Analysis

4.8.1 Descriptive Analysis of the NGO Network

We first extract information on NGO headquarters from the iCSO database and examine whether countries in the global North are overrepresented in the network, as argued in Hypothesis 1. Among the 1200 NGOs in our sample, 1, 116 have addresses listed in iCSO; 102 different countries are listed as the address for these organizations. Figure 4.5 shows the distribution of the top 10 countries of residence of the NGOs in our sample.

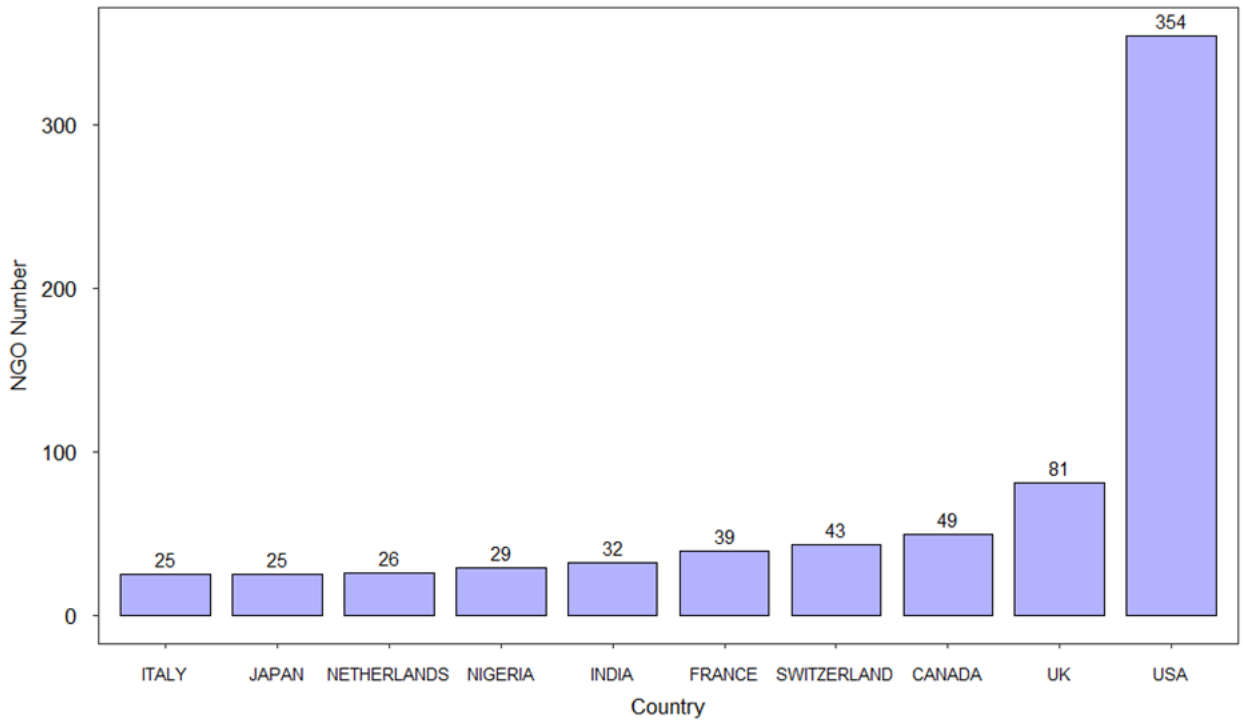


Figure 4.5: Top 10 country distribution of NGOs in NGO network

There is no universal list or definition of countries in the global North. For our study, we code addresses of organizations in the 36 countries that are members of the Organization of Economic Co-operation and Development (OECD) as global North organizations; all other organizations with addresses listed are categorized as organizations from the global South. As Figure 4.6 shows, despite much discussion of the issue of power imbalances in the NGO network, even when focusing just on 2002-2017, after this problem had been widely discussed in the network-as-structure and practitioner literature, we see very little movement in the percentage of global South NGOs involved in the network over time.

Figure 4.6: Comparing global south and global north NGO percentages in the network

To test Hypothesis 1, we focus on differences between the average degree centrality scores between organizations from the global North and the global South. Degree is a basic network measure of the total number of ties connecting an actor to others in the total network (Wasserman, Faust et al., 1994). Figure 4.7 provides an illustration of the distribution of degree over time. Again, to be as stringent as possible, we focus this test only on the years 2002-2017, after there had been widespread discussion of the power disparities between organizations in the global North and global South. Results of Wilcoxon rank sum test ($p < 0.05$) in each year support Hypothesis 1: greater participation in the NGO network is associated with global North status.

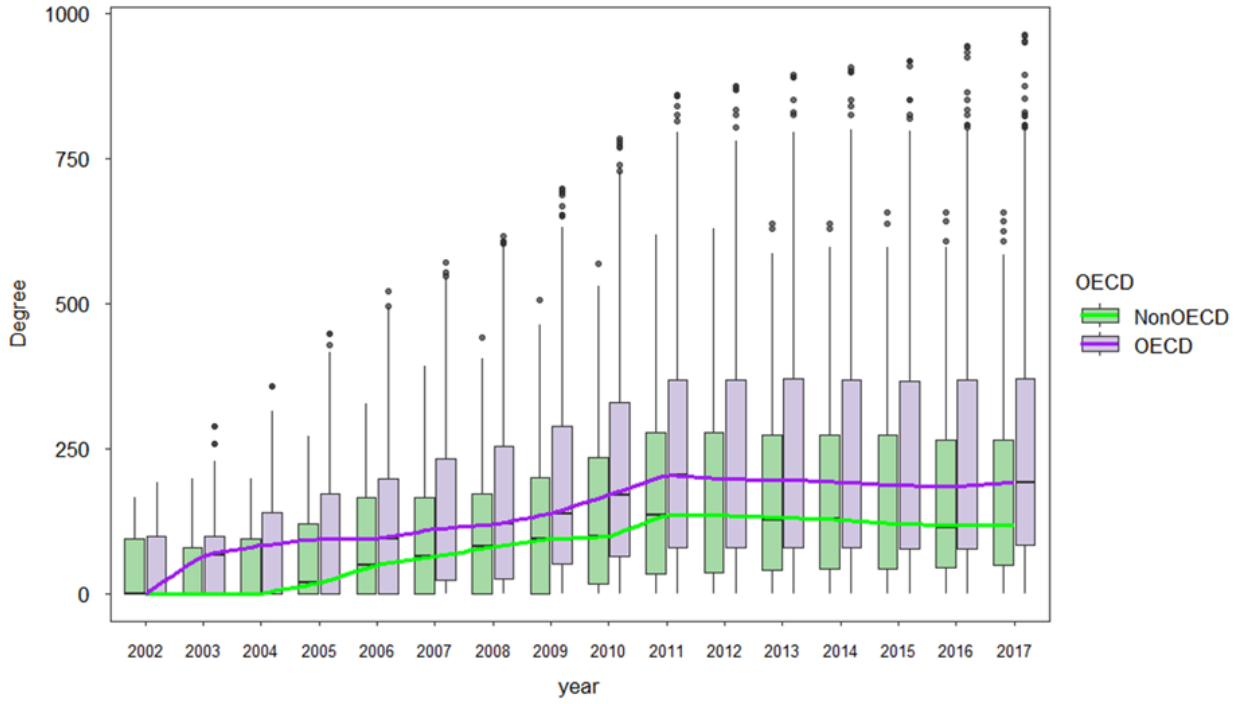


Figure 4.7: Average degree centrality over time within the NGO network

Another measure of centrality, betweenness centrality, closely relates to ideas of brokerage. In fact, it has been referred to as "brokerage capacity" in extant literature (Caniglia, 2001). Betweenness centrality is thought to capture those that are in the "middle" or "bridges" in that it captures the total number of shortest paths between nodes that go through a particular node (A. Murdie & Davis, 2012; Wasserman, Faust et al., 1994). Figure 4.8 provides the OECD and non-OECD distribution of betweenness centrality scores over time. Results of the Wilcoxon rank test sum ($p < 0.05$) in each year also support Hypothesis 1: global North organizations are more likely to participate more in the network, even in ways that resemble brokerage.

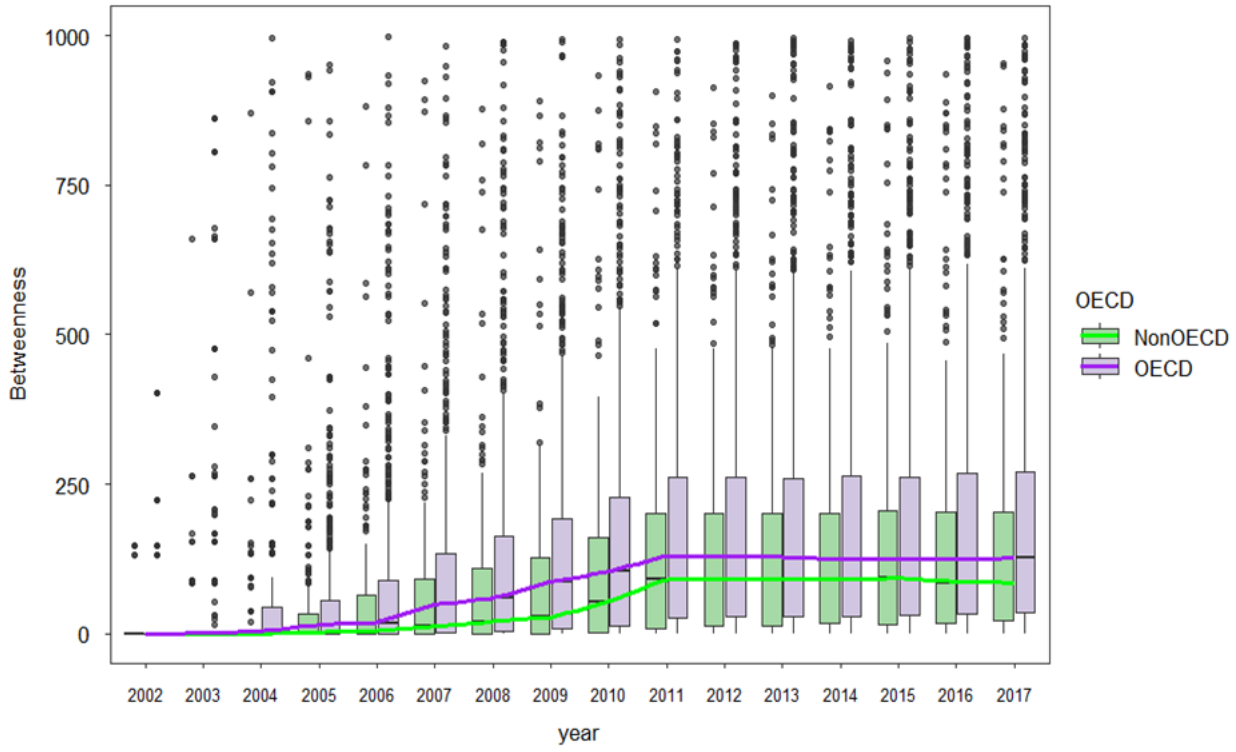


Figure 4.8: Average betweenness centrality over time within the NGO network

4.8.2 Community Detection

Our next task was to detect emerging communities within the NGO network. We used a time-varying stochastic block model to identify the optimal common composition of communities that maximizes modularity across years. For the static network at time t, the concept of modularity was introduced for networking clustering in 2004 (M. E. Newman & Girvan, 2004)

$$modularity(t) = \frac{1}{2m_t} \sum_{\text{i,j in the same community}} [A_{ij}(t) - \frac{k_i(t)k_j(t)}{2m_t}]$$

Where $A_{ij}(t)$ is the number of edges between vertices i and j (the quantities $A_{ij}(t)$ are the elements of the so-called adjacency matrix), $k_i(t)$ and $k_j(t)$ are the degrees of the vertices i and j, m_t is the total number of edges in the network $m_t = \frac{1}{2}\sum_i k_i(t)$, and the leading factor $\frac{1}{2m_t}$ of is merely conventional

for normalization. Zhang and Cao (J. Zhang & Cao, 2017) defines the mean modularity of different time as the final modularity of the time-varying network. Modularity varies from 0 to 1. A higher modularity value indicates denser connections between nodes within communities (groups) and sparser connections between communities (Cheng et al., 2018; M. E. Newman & Girvan, 2004; J. Zhang & Cao, 2017). In this chapter, we adopt the Louvian maximization method, which is a greedy community detection method (Blondel et al., 2008). The algorithm iteratively builds new communities until the modularity reaches its maximum. Thus, the algorithm automatically provides us a rationale for the number of communities identified.

As shown in Table 4.1, the four identified communities have between 186 and 354 organizations each. Community 3 has the highest mean degree centrality score for its organizations; however, it also has the lowest average betweenness centrality score. Community 4 has the highest mean betweenness centrality. To note, betweenness centrality and degree here record ties to the whole network; our brokerage role analysis, discussed below, discusses how ties differ between and within communities. The modularity value (0.25) of our network suggests many cross-community ties, providing opportunities for brokers.

Table 4.1: Community detection

Community Label	Community 1	Community 2	Community 3	Community 4
Nodes count	350	310	186	354
Degree Average	143.87	317.32	329.71	167.27
Betweenness Average	550.48	535.89	430.04	628.63

How can we explain the character of these emergent communities? Word clouds of the names of the organizations are shown in Figure 4.9. These help us identify some interesting patterns in the communities. First, Community 1 is composed of many smaller organizations that have names that reflect geographic locations and regions. There appears to be a focus on indigenous or people's rights.

Community 2 is comprised of what many would consider the classic or textbook international NGOs. There are quite a few organizations in Community 2 that are household names: Amnesty International, CARE International, Open Society, and Oxfam, for example. Although there are many mentions of women or women's rights, there are many other organizations that appear generally focused on human rights and development.

Community 3 includes many professional organizations, like the International Sociological Association, the American Psychological Association, and the International Studies Association. It also includes many research-focused foundations at both ends of the political spectrum, like the Kroc Institute for Peace and Justice, Focus on the Family, the Heritage Foundation, and the International Planned Parenthood Federation. The final community, Community 4, appears to have a large proportion of environmental organizations, including Greenpeace International, Earth Action, Friends of the Earth International, and the Sierra Club.

The composition of the communities detected has a lot of face validity for NGO scholars. Communities 1 and 4 appear to have some issue that may be helping to drive the structure. Community 2 is comprised by many high-status organizations; it is fascinating that the algorithm was able to group so many of these together with no other information other than their network data. Finally, the ideological diversity in Community 3 is especially interesting. In some regards, these organizations are working with very separate and competing agendas.

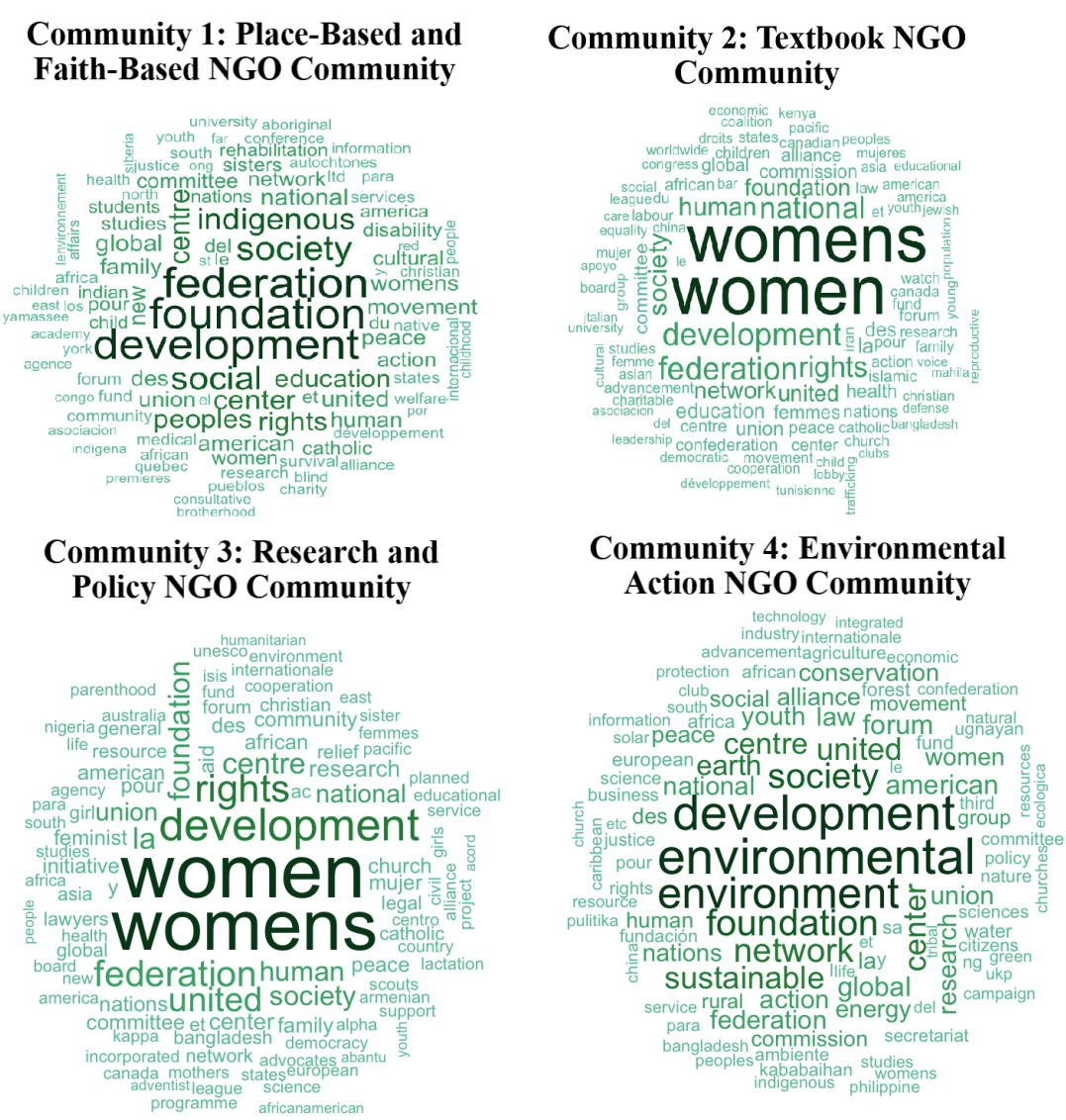


Figure 4.9: Community grouping

4.8.3 Brokerage Roles

Hypothesis 2 concerns the distribution of brokerage roles across communities. As we argued, some communities are likely to have more coordinators, connecting organizations mainly within a specific community, while other communities could have more itinerants, gatekeepers/representatives, or liais-

ons, connecting with organizations outside of their specific community in divergent ways. In this way, brokerage provides power to organizations differently across the communities.

Brokers must connect two organizations. As such, only organizations with at least two neighbors were assigned a specific role. Given our communities, brokerage properties were determined in line with Gould and Fernandez (Gould & Fernandez, 1989b) using the "brokerage" command in the "sna" R package (Butts et al., 2008). Table 4.2 provides a breakdown of the brokerage role distribution in each community. A Chi-squared test allows us to reject the null hypothesis that there is no association between community and brokerage role at the $p < 0.05$ level, providing support to Hypothesis 2. As expected, there is an association between community and brokerage role.

Table 4.2: Brokerage role distribution by community

Role	Community 1	Community 2	Community 3	Community 4
Coordinator	0.55	0.43	0.00	0.70
Gatekeeper Representative	0.35	0.50	0.49	0.23
Liaison	0.10	0.07	0.29	0.07
Itinerant	0.00	0.00	0.22	0.00

Although the Chi-squared test supported Hypothesis 2 , when taken with a qualitative understanding of the organizations in each community, there are many things in Table 4.2 that are both interesting and surprising. First, Communities 1 and 4 are similar in that they have their largest percentages of brokers in coordinator roles, connecting unconnected organizations within their communities. For both communities, over half of their brokers are coordinators, operating inside the community, instead of outside the community. Community 4, where we identified many environmental organizations, has the highest percentage of brokers (70%) that are coordinators. Perhaps this issue area is particularly prone to isolation from the overall network; organizations in Community 4 may see limited value in brokering with other communities.

Community 2 was the community with many of the household-name organizations. Although there is still a large percentage of coordinators, indicating that brokers in this community still work to connect others within the community, half of the brokerage roles are gatekeeper/representative roles. This is consistent with much of the "network-as-structure" critiques of these household-name organizations (Bob, 2005; Carpenter, 2014). Somewhat surprisingly, there are no itinerant brokers in this group and a very low percentage of liaison brokers (7%). There are many organizations in Community 2 that Stroup and Wong (Stroup & Wong, 2017) classified as "leading" organizations. However, when compared to Community 3, where there is a higher percentage of liaison and itinerant brokers, Community 2 organizations are somewhat less moderate in their advocacy strategy and may have less need to connect organizations in another community or to connect organizations across communities.

Community 3, where we saw many professional organizations and partisan research organizations, is perhaps the most interesting and divergent in its role distribution. It is the only community to have no coordinators, indicating that brokerage between organizations in this community may be of little value to the organizations. Given the divergent partisan bent of many of the organizations in this community, this could be expected. Community 3 is also the only community where we find itinerant brokers. This could be linked to the large percentage of professional organizations in this community; these organizations often have missions to support the professional interests of their members to the outside world.

Given this discussion of the various distribution of roles across these communities, we now examine the distribution of global North organizations within specific brokerage roles. Within the specific communities, our third hypothesis was that brokerage roles are associated with global North status. Global North organizations may be more likely to get those roles that broker relationships across communities, namely gatekeepers/representatives, liaisons, and itinerant brokers.

It is worth noting that there is no association ($p > 0.05$) between community and OECD status: all communities are made up of between 67% and 74% OECD organizations. In all four communities, however, we do find an association between brokerage role types and OECD status ($p < 0.05$). Figure

4.10 summarizes the percentage of OECD organizations in each brokerage role in each community. The horizontal line in each graph signifies the overall percentage of OECD organizations in each community.

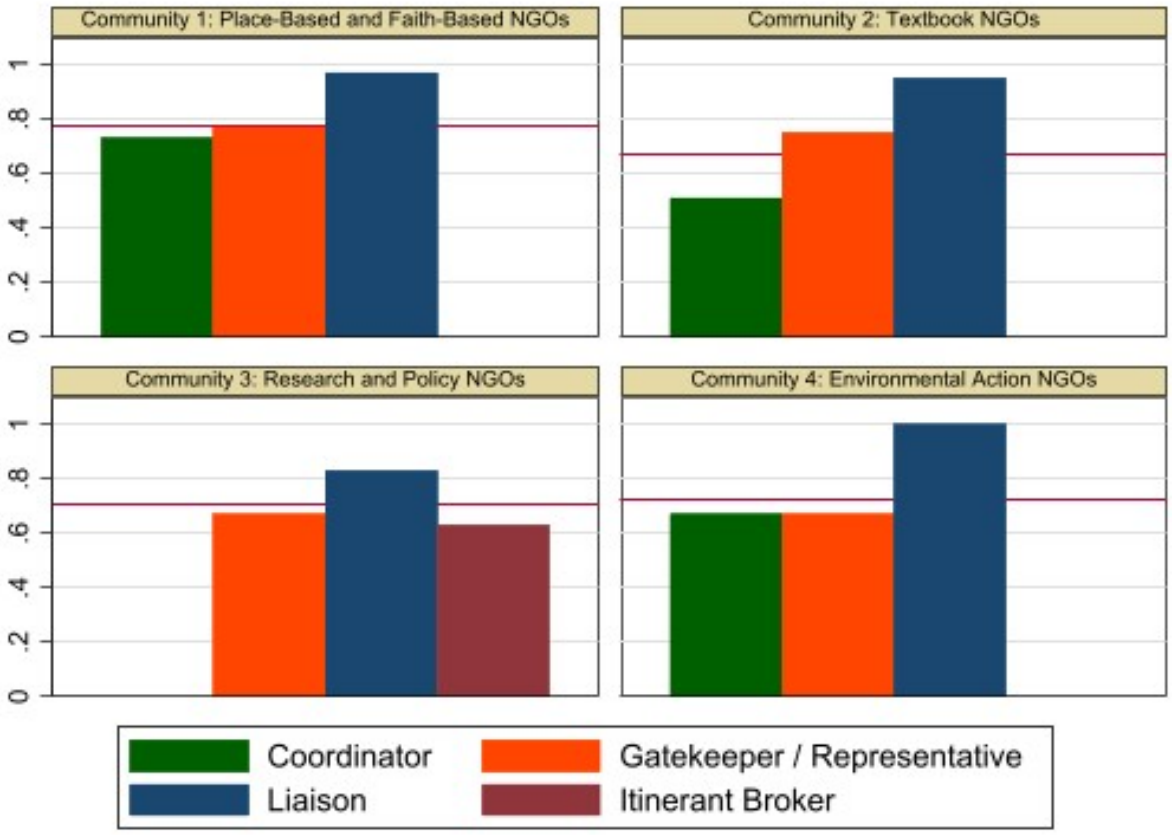


Figure 4.10: OECD status and role

In the communities where there are coordinators (Communities 1, 2, and 4), these coordinator roles are filled with the lowest percentages of global North organizations. Gatekeepers and liaison roles are filled with a higher percentage of global North organizations. This is consistent with our expectations and with the extant literature: organizations with pre-existing power have been able to gain additional power through the advocacy network (Bob, 2005). Their brokerage roles could help reinforce their power within their community as a gatekeeper/representative and between multiple communities as a liaison.

As before, Community 3, the community with many partisan and/or professional groups, shows a very different pattern. Since this community is likely to contain organizations whose goals may include reaching outside of professional or issue boundaries, it is somewhat expected that this group has an increased percentage of global South organizations in these brokerage roles. The results concerning itinerant brokers were surprising to us: only 61% of itinerant brokers are from the global North. Perhaps, for organizations in this particular community, networking into a separate community is a special priority, making it the brokerage strategy for both organizations from the global North and the global South. If the organization had a little prior status in Community 3, however, this brokerage role may provide limited status gains in the organization's community (Sullivan & Stewart, 2017).

The social network analysis of this new dataset not only provides support for our hypotheses, it also provides many insights into the evolving structure of networked advocacy and the current NGO environment. The transnational advocacy network may be a public good, increasing perceived success and output (A. Murdie, 2014; Tallberg, Dellmuth et al., 2018). It is also an unequal power environment, with comparatively little participation from the majority of the world's population living outside of the global North.

Further, the network is evolving into distinct communities, with human rights and development organizations that are household names being in one community (Community 2), separate from communities that focus on environmental issues (Community 4), indigenous or regional issues (Community 1), and professional and partisan research organizations (Community 3). Professional and partisan research organizations, although long recognized as part of the INGO community (Boli & Thomas, 1999), have received comparatively little attention in the NGO literature, especially from within international relations. Our examination of community and brokerage, however, shows how different organizations in this community (Community 3) are in their networking behavior from the rest of the communities in our network. This community does not have coordinators and was the only community where we identified itinerant brokers. Further research on this particular community and how it interacts with other NGO communities is definitely necessary.

Our analysis of brokerage roles also provides many insights into the workings of the NGO network and how pre-existing power differences may be associated with certain brokerage roles within communities. In the communities comprised of organizations that have been the traditional focus of international relations (Communities 1, 2, and 4), global North organizations are not only more likely to be involved in the network, they are more likely to take liaison and gatekeeper/representative roles within their community These roles enable organizations to shape the advocacy agenda their community shares with the outside world, ultimately reinforcing and growing their personal power and status. For each community that represents the traditional focus of international relations (Communities 1, 2, and 4), a larger percentage of organizations from the global South are in a coordinator broker role. Although this role may still give these organizations access to the network, a coordinator role will not provide the agenda-setting or resource-allocating power associated with gatekeeper/representative or liaison roles. Through evolving communities and disparities in roles, the advocacy network could be exacerbating power inequalities.

4.9 Conclusion

Like a pointillist painting, understanding TANs and NGOto-NGO networking requires us to gaze at the "complexity" of the overall network structure while also closely examining the "dots" or NGOs that make up that structure . In this project, we have jointly examined the complex community structure of the overall network and the distribution of distinct brokerage roles within the network, providing a richer, more complete picture of TANs and NGO networking. Our work provides insights into a long-standing puzzle in the NGO literature: namely, why the "network-as-actor" literature sees so much promise in networks while the "network-as-structure" literature sees so much exploitation. While networking can create communities that help solve problems and share knowledge, the broker NGOs that connect these communities may be able to use their specific brokerage positions to garner more power for themselves.

A renewed emphasis on community detection and communities of practice insights can provide further insights on NGOs and TANs. We need to move beyond exogenously assuming that organizations within the same issue area network together. Instead, NGOs have problems and interests that lead more

diverse communities to emerge. Some of these communities extend beyond ideological or partisan divides (Community 3: Research and Policy NGO Community), while other communities may be dominated by leading NGOs (Community 2: Textbook NGO Community) or certain coalitions of causes (Community 1: Place-Based and Faith-Based NGO Community and Community 4: Environmental Action NGO Community). Community detection methods help us establish the boundaries between communities of practice, eliminating the "vagueness" previously identified in communitarian approaches. Further, by disaggregating brokerage into distinct brokerage roles, we identify variation in NGO and community networking practices. Some emergent NGO communities develop practices that privilege brokerage roles that connect organizations across communities, while others may comprise brokers that focus only within their community, potentially siloing off their community's advocacy causes and practices. Brokerage may help innovation, but it also reinforces hierarchy and power disparities. Preexisting power disparities, like between NGOs from the global North and the global South, lead to differences in brokerage role distributions. To our knowledge, the disaggregation of brokerage into distinct roles is not widespread in IR but has much potential for further theorizing both inside and outside of the study of TANs and NGOs. Finally, our study offers a cautionary note for NGO practitioners and UN officials interested in ensuring a multitude of NGO voices at UN-facilitated meetings. Even if current efforts to increase access to global South NGOs are successful, brokerage role dynamics and the community structure of the overall advocacy network may lead global South NGOs to be relegated to certain less powerful brokerage roles within communities, as opposed to the more powerful brokerage roles that transmit information and resources across communities. Hierarchy and power dynamics persist and could even be heightened as a result of wellintentioned but poorly planned initiatives to increase global South involvement. Our dataset and findings could help in improving initiatives to increase global South representation in the NGO network, especially at UN meetings. First, future studies could use our dataset and findings to identify global South NGOs that are well embedded in the NGO network and occupy powerful brokerage roles between communities.17 Researchers and practitioners could examine these cases to identify common characteristics that help global South organizations flourish in the current NGO network. Examining

these characteristics can provide us with potential best practices and may assist concerned donors with identifying organizations that are well positioned or "ripe" for additional assistance. Second, our results suggest that efforts to increase global South involvement and representation across communities may be more helpful than efforts that facilitate involvement within a specific community. This may require grants or donations to foster attendance at multiple NGO meetings, hoping to facilitate connections in multiple communities of practice. To the extent that voices from global South NGOs are critical to UN legitimacy, easing the ability for NGOs to be involved in multiple meetings may help improve public opinion about the UN. Moreover, simply acknowledging that historically powerful actors still dominate the NGO network could help encourage organizational reflection. In their attempt to build connections between organizations, those NGOs with preexisting power may get additional power. If global North organizations are committed to improved representation of global South NGOs, they may need to avoid brokering connections between other NGOs and instead create spaces where NGO ties can develop directly. Through these changes, the advocacy network may be better equipped to listen and respond to the plight of the world's powerless in ways that do not reinforce and exacerbate power disparities.

Appendix A

Appendix for Chapter 2

This appendix is organized as follows.

- Section A.1 presents notation table.
- Section A.2 provides proofs of all theoretical results.
- Section A.3 provides additional simulation results.

A.1 Notation Table

Table A.1 summarizes the notations in the main text.

O_p, o_p Notation. For a sequence of random variables, $\{A_n\}$, and a sequence of constants, $\{a_n\}$, the notation $A_n = O_p(a_n)$ means that $\{A_n/a_n\}$ is stochastically bounded (or bounded in probability). That is, for any $\tau > 0$, there exists a constant $\kappa(\tau)$ and an integer $n(\tau)$ such that if $n \geq n(\tau)$, then

$$P(|A_n/a_n| \leq \kappa(\tau)) \geq 1 - \tau.$$

Table A.1: Notation table

Notation	Interpretation
n	Number of nodes
$\mathbf{A} = (a_{ij})$	Adjacency matrix
$\tilde{\mathbf{A}}^{(k)} = (\tilde{a}_{ij}^{(k)})$	kth masked adjacency matrix
$\mathbf{P}_0 = (p_{ij})$	True probability matrix
M	Candidate hyperparameter
m	Number of candidates
$\mathcal{M}$	Set of candidates, $\{M_1, \ldots, M_m\}$
$\mathbf{P}_M^{\mathbf{A}} = (\hat{p}_{ij})$	Estimate using hyperparameter M and data $\mathbf{A}$
$\overset{*}{\mathbf{P}}{}_M^{\tilde{\mathbf{A}}^{(k)}} = (\overset{*}{p}{}_{ij}^{(k)})$	Debiased estimate using M and $\tilde{\mathbf{A}}^{(k)}$
$\mathbb{P}(\cdot)$	Probability measure
$\text{MSE}(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0)$	MSE between true and estimator
$\text{MMV}_M(\mathbf{A})$	MMV score of M for data $\mathbf{A}$
$\mathcal{S}_{\mathbf{A}}$	Set of all node pairs in $\mathbf{A}$
K	Number of folds
$\mathcal{S}_k$	kth set of node pairs, used for validation
$\mathcal{S}_{-k}$	The rest $k-1$ sets of node pairs, $S_{\mathbf{A}}/S_k$, for training
θ	A parameter deciding the distribution of $\tilde{B}$
λ	A parameter deciding K, where $K = \lambda n$
M_o^{mse}	Optimal hyperparameter selected by MSE
M_o	Optimal hyperparameter selected by MMV

The notation $A_n = o_p(a_n)$, means that $\{A_n/a_n\}$ converges to zero in probability. Equivalently, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|A_n/a_n| > \epsilon) = 0.$$

The notation $A_n = \Omega_p(a_n)$ means that $\{A_n/a_n\}$ is stochastically lower bounded (or bounded in probability). That is, for any $\tau > 0$, there exists a constant $\kappa(\tau)$ and an integer $n(\tau)$ such that if $n \geq n(\tau)$, then

$$P(|A_n/a_n| \geq \kappa(\tau)) \geq 1 - \tau.$$

The notation $A_n = \Theta_p(a_n)$ means that there exists a constant τ such that

$$\lim_{n\to\infty} P(|A_n/a_n| = \tau) = 1.$$

Remark. If $Var(A_n) = O(n^{2\delta})$ and $E(A_n) = 0$, where δ is a real number, then we have that $\{A_n/n^{\delta}\}$ is bounded in probability by Chebyshev's inequality. We write $A_n = O_p(n^{\delta})$.

A.2 Technical Proofs

A.2.1 Proof of Theorem 1

Let $\tilde{\mathbf{A}}^{(k)}$ denote the kth fold training graph. We define

$$V\left(\overset{*}{\mathbf{P}}{}_M^{\tilde{\mathbf{A}}^{(k)}}, \mathbf{P}_0\right) = \frac{1}{|\mathcal{S}_k|} \sum_{\substack{v_i v_j \in \mathcal{S}_k \\ i<j}} (\overset{*}{p}{}_{ij}^{(k)} - p_{ij})^2$$

Note that we are doing $K-$fold cross-validation. The sets $\mathcal{S}_1, \cdots \mathcal{S}_K$ are disjoint. With a little abuse of the notation, we write $\overset{*}{p}_{ij} = \overset{*}{p}{}_{ij}^{(k)}$ for all $i < j, v_i v_j \in \mathcal{S}_k$ and $k = 1, \cdots, K$. We then define

$$\begin{aligned} V_M(\mathbf{P}_0) &= \frac{1}{K}\sum_{k=1}^{K} V\left(\overset{*}{\mathbf{P}}{}_M^{\tilde{\mathbf{A}}^{(k)}}, \mathbf{P}_0\right) \\ &= \frac{1}{n(n-1)/2} \sum_{1\le i<j\le n} (\overset{*}{p}_{ij} - p_{ij})^2 \end{aligned}$$

To prove Theorem 1, it suffices to show that

$$\mathrm{MMV}_M(\mathbf{A}) - V_M(\mathbf{P}_0) = O_p(\frac{1}{n}) \tag{A.1}$$

$$V_M(\mathbf{P}_0) - MSE\left(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0\right) = O_p(\frac{1}{K}) \tag{A.2}$$

We first prove (A.1). By equations (9) and (10), we have

$$\begin{aligned}\mathrm{MMV}_M(\mathbf{A}) - V_M(\mathbf{P}_0) &= \frac{1}{n(n-1)/2}\sum_{1\le i<j\le n}\left((\overset{*}{p}_{ij}-a_{ij})^2-(\overset{*}{p}_{ij}-p_{ij})^2\right)\\ &= \frac{1}{n(n-1)/2}\sum_{1\le i<j\le n}\left((p_{ij}-a_{ij})^2+(\overset{*}{p}_{ij}-p_{ij})(p_{ij}-a_{ij})\right)\end{aligned}$$

Notice that a_{ij} is independent from $\tilde{\mathbf{A}}_k$ for all $i<j, v_iv_j\in\mathcal{S}_k$. Thus a_{ij} is also independent of $\overset{*}{p}_{ij}$. We then have

$$\begin{aligned}\mathbb{E}(\mathrm{MMV}_M(\mathbf{A}) - V_M(\mathbf{P}_0) - V) &= \frac{1}{n(n-1)/2}\sum_{1\le i<j\le n}\left(\mathbb{E}(p_{ij}-a_{ij})^2+\mathbb{E}(\overset{*}{p}_{ij}-p_{ij})\mathbb{E}(p_{ij}-a_{ij})\right)-V\\ &= \frac{1}{n(n-1)/2}\sum_{1\le i<j\le n}\left(\mathbb{E}(p_{ij}-a_{ij})^2\right)-V\\ &= \frac{1}{n(n-1)/2}\sum_{1\le i<j\le n}p_{ij}(1-p_{ij})-V\\ &= 0. \end{aligned}\tag{A.3}$$

To derive the order of $\mathrm{MMV}_M(\mathbf{A})-V_M(\mathbf{P}_0)-V$, we need to derive the order of $\mathrm{Var}(\mathrm{MMV}_M(\mathbf{A})-V_M(\mathbf{P}_0)-V)$, which can be decomposed as follows

$$\begin{aligned}&\mathrm{Var}(\mathrm{MMV}_M(\mathbf{A}) - V_M(\mathbf{P}_0) - V) = \mathrm{Var}(\mathrm{MMV}_M(\mathbf{A}) - V_M(\mathbf{P}_0))\\ =\ &(\frac{1}{n(n-1)/2})^2\sum_{1\le i<j\le n}\{\mathrm{Var}\left((p_{ij}-a_{ij})(\overset{*}{p}_{ij}-a_{ij})\right)\\ +\ &\sum_{1\le i<j\le n}\sum_{\substack{1\le q<r\le n\\ q\ne i \text{ or } r\ne j}}\mathrm{Cov}\left((p_{ij}-a_{ij})(\overset{*}{p}_{ij}-a_{ij})\ ,(p_{qr}-a_{qr})(\overset{*}{p}_{qr}-a_{qr})\right)\}.\end{aligned}\tag{A.4}$$

Notice that in the first term,

$$\begin{aligned}\mathrm{Var}\left((p_{ij}-a_{ij})(\overset{*}{p}_{ij}-a_{ij})\right) &= \mathbb{E}\left((p_{ij}-a_{ij})(\overset{*}{p}_{ij}-a_{ij})-p_{ij}(1-p_{ij})\right)^2\\ &= p_{ij}\mathbb{E}[(p_{ij}-1)(\overset{*}{p}_{ij}-1)-p_{ij}(1-p_{ij})]^2\\ &+ (1-p_{ij})\mathbb{E}[p_{ij}\overset{*}{p}_{ij}-p_{ij}(1-p_{ij})]^2\\ &= p_{ij}(1-p_{ij})\mathbb{E}(1-\overset{*}{p}_{ij}-p_{ij})^2\end{aligned}\tag{A.5}$$

It is easy to see that there exists a global constant Γ such that

$$p_{ij}(1-p_{ij})\mathbb{E}(1-\overset{*}{p}_{ij}-p_{ij})^2<\Gamma,$$

we thus have

$$(\frac{1}{n(n-1)/2})^2\sum_{1\le i<j\le n}\text{Var}\left((p_{ij}-a_{ij})(\overset{*}{p}_{ij}-a_{ij})\right)=O(\frac{1}{n^2}) \tag{A.6}$$

We then derive the order of the second term in (A.4). To ease the description, we define $\eta^0_{ij,qr}\overset{d}{=}(1-p_{ij}-\overset{*}{p}_{ij})|a_{qr}=1,\eta^1_{ij,qr}\overset{d}{=}(1-p_{ij}-\overset{*}{p}_{ij})|a_{qr}=0$.

According to the law of total expectation, we have

$$\begin{aligned}
&\text{Cov}\left((p_{ij}-a_{ij})(\overset{*}{p}_{ij}-a_{ij}),(p_{qr}-a_{qr})(\overset{*}{p}_{qr}-a_{qr})\right)\\
=\;&((1-p_{ij}-\overset{*}{p}_{ij})(1-p_{qr}-\overset{*}{p}_{qr})(a_{ij}-p_{ij})(a_{qr}-p_{qr})\\
=\;&\mathbb{P}(a_{ij}=1,a_{qr}=1)(1-p_{ij})(1-p_{qr})\mathbb{E}((1-p_{ij}-\overset{*}{p}_{ij})(1-p_{qr}-\overset{*}{p}_{qr})|a_{ij}=1,a_{qr}=1)\\
&-\mathbb{P}(a_{ij}=1,a_{qr}=0)(1-p_{ij})p_{qr}\mathbb{E}((1-p_{ij}-\overset{*}{p}_{ij})(1-p_{qr}-\overset{*}{p}_{qr})|a_{ij}=1,a_{qr}=0)\\
&-\mathbb{P}(a_{ij}=0,a_{qr}=1)p_{ij}(1-p_{qr})\mathbb{E}((1-p_{ij}-\overset{*}{p}_{ij})(1-p_{qr}-\overset{*}{p}_{qr})|a_{ij}=0,a_{qr}=1)\\
&+\mathbb{P}(a_{ij}=0,a_{qr}=0)p_{ij}p_{qr}\mathbb{E}((1-p_{ij}-\overset{*}{p}_{ij})(1-p_{qr}-\overset{*}{p}_{qr})|a_{ij}=0,a_{qr}=0)\\
=\;&p_{ij}p_{qr}(1-p_{ij})(1-p_{qr})\left\{\mathbb{E}(\eta^1_{ij,qr}\eta^1_{qr,ij})-\mathbb{E}(\eta^1_{ij,qr}\eta^0_{qr,ij})-\mathbb{E}(\eta^0_{ij,qr}\eta^1_{qr,ij})+\mathbb{E}(\eta^0_{ij,qr}\eta^0_{qr,ij})\right\}\\
=\;&p_{ij}p_{qr}(1-p_{ij})(1-p_{qr})\mathbb{E}\left((\eta^1_{ij,qr}-\eta^0_{ij,qr})(\eta^1_{qr,ij}-\eta^0_{qr,ij})\right),
\end{aligned} \tag{A.7}$$

where the expectation in (A.7) indicates taking expectation with respect to all the random elements except a_{ij} and a_{qr}. By using Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\left|\text{Cov}\left((p_{ij}-a_{ij})(\overset{*}{p}_{ij}-a_{ij}),(p_{qr}-a_{qr})(\overset{*}{p}_{qr}-a_{qr})\right)\right|^2\\
\le\;&\mathbb{E}\left(p_{ij}(1-p_{ij})(\eta^1_{ij,qr}-\eta^0_{ij,qr})\right)^2\mathbb{E}\left(p_{qr}(1-p_{qr})(\eta^1_{qr,ij}-\eta^0_{qr,ij})\right)^2.
\end{aligned} \tag{A.8}$$

By combining the results of equation (A.7), and (A.8) we have

$$
\begin{aligned}
&\sum_{1\le i<j\le n}\sum_{1\le q<r\le n}\mathrm{Cov}\left((p_{ij}-a_{ij})(\overset{*}{p}_{ij}-a_{ij}),(p_{qr}-a_{qr})(\overset{*}{p}_{qr}-a_{qr})\right)\\
\le\ &\sum_{\substack{1\le i<j\le n\\1\le q<r\le n}}\left(\mathbb{E}\left(p_{ij}(1-p_{ij})(\eta^1_{ij,qr}-\eta^0_{ij,qr})\right)^2\mathbb{E}\left(p_{qr}(1-p_{qr})(\eta^1_{qr,ij}-\eta^0_{qr,ij})\right)^2\right)^{\frac{1}{2}}\\
\le\ &\frac{1}{16}\sum_{\substack{1\le i<j\le n\\1\le q<r\le n}}\left(\mathbb{E}(\eta^1_{ij,qr}-\eta^0_{ij,qr})^2\mathbb{E}(\eta^1_{qr,ij}-\eta^0_{qr,ij})^2\right)^{\frac{1}{2}}\\
\le\ &\frac{1}{16}\sum_{1\le i<j\le n}\mathbb{E}(\eta^1_{ij,qr}-\eta^0_{ij,qr})^2 \qquad \text{(A.9)}
\end{aligned}
$$

where the inequality (A.9) is by Cauchy-Schwarz inequality.

Then, we verify the value of $\mathbb{E}(\eta^1_{ij,qr}-\eta^0_{ij,qr})^2$. Suppose $a_{ij}\in\mathcal{S}_k$. If $a_{qr}\in\mathcal{S}_k$ also, the training graph $\tilde{\mathbf{A}}^{(k)}$ is independent from a_{qr}. Thus, $\eta^1_{ij,qr}\overset{d}{=}\eta^0_{ij,qr}$. In this case, we can see that $\mathbb{E}(\eta^1_{ij,qr}-\eta^0_{ij,qr})^2=0$.

If $a_{qr}\notin\mathcal{S}_k$, let $\tilde{\mathbf{A}}^{(k)}_0$ denote the training graph when $a_{qr}=0$ and $\tilde{\mathbf{A}}^{(k)}_1$ denote the training graph when $a_{qr}=1$. All the other entries except a_{qr} are the same for $\tilde{\mathbf{A}}^{(k)}_0$ and $\tilde{\mathbf{A}}^{(k)}_1$. Under Assumption 2, a global constant C exists such that,

$$
\begin{aligned}
\sum_{1\le i<j\le n}\mathbb{E}(\eta^1_{ij,qr}-\eta^0_{ij,qr})^2 &\le \mathbb{E}\|\overset{*}{\mathbf{P}}{}^{\tilde{\mathbf{A}}^{(k)}_1}_M-\overset{*}{\mathbf{P}}{}^{\tilde{\mathbf{A}}^{(k)}_0}_M\|_F^2\\
&\le C^2\mathbb{E}\|\tilde{\mathbf{A}}^{(k)}_1-\tilde{\mathbf{A}}^{(k)}_0\|_F^2\\
&= C^2. \qquad \text{(A.10)}
\end{aligned}
$$

Combining (A.7) (A.9), (A.10), we get

$$
\begin{aligned}
&\left(\frac{1}{n(n-1)/2}\right)^2\sum_{1\le i<j\le n}\sum_{1\le q<r\le n}\mathrm{Cov}\left((p_{ij}-a_{ij})(\overset{*}{p}_{ij}-a_{ij}),(p_{qr}-a_{qr})(\overset{*}{p}_{qr}-a_{qr})\right)\\
\le\ &\left(\frac{1}{n(n-1)/2}\right)C^2\\
=\ &O(\frac{1}{n^2}) \qquad \text{(A.11)}
\end{aligned}
$$

Thus, combining (A.3), (A.4),(A.6) and (A.11), we have

$$\mathrm{MMV}_M(\mathbf{A}) - V_M(\mathbf{P}_0) = O_p(\frac{1}{n}), \tag{A.12}$$

by Chebyshev's inequality.

We then prove that (A.2) holds.

$$\begin{aligned}
|V\left(\overset{*}{\mathbf{P}}{}_M^{\tilde{\mathbf{A}}^{(k)}}, \mathbf{P}_0\right) - MSE\left(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0\right)| &= \left|\frac{1}{n(n-1)/2}\sum_{1\le i<j\le n}(\hat{p}_{ij}^* - \hat{p}_{ij})^2 - (\hat{p}_{ij} - p_{ij})^2\right| \\
&\le \frac{1}{n(n-1)/2}\sum_{1\le i<j\le n}\left|(\hat{p}_{ij}^* - \hat{p}_{ij})^2 - (\hat{p}_{ij} - p_{ij})^2\right| \\
&\le \frac{1}{n(n-1)}\|\mathbf{P}_M^{\mathbf{A}} - \overset{*}{\mathbf{P}}{}_M^{\tilde{\mathbf{A}}^{(k)}}\|_F^2 \\
&= \frac{1}{n^2}\|\mathbf{P}_M^{\mathbf{A}} - \frac{K}{K-1}\mathbf{P}_M^{\tilde{\mathbf{A}}^{(k)}} + \frac{\theta}{K-1}\|_F^2 \\
&\le \frac{1}{n^2}\{(\frac{K}{K-1})^2\|\mathbf{P}_M^{\mathbf{A}} - \mathbf{P}_M^{\tilde{\mathbf{A}}}\|_F^2 \\
&+ (\frac{1}{K-1})^2\|\mathbf{P}_M^{\tilde{\mathbf{A}}^{(k)}} - \theta\|_F^2\},
\end{aligned} \tag{A.13}$$

Since that $\|\mathbf{A} - \tilde{\mathbf{A}}^{(k)}\|_F^2 = O_p(\frac{n^2}{K})$. Under Assumption 2, we have

$$\begin{aligned}
\|\mathbf{P}_M^{\tilde{\mathbf{A}}^{(k)}} - \mathbf{P}_M^{\mathbf{A}}\|_F^2 &= \frac{\|\mathbf{P}_M^{\tilde{\mathbf{A}}^{(k)}} - \mathbf{P}_M^{\mathbf{A}}\|_F^2}{\|\mathbf{A} - \tilde{\mathbf{A}}^{(k)}\|_F^2}\|\mathbf{A} - \tilde{\mathbf{A}}^{(k)}\|_F^2 \\
&= O_p(\frac{n^2}{K}).
\end{aligned} \tag{A.14}$$

Meanwhile, we have

$$(\frac{1}{K-1})^2\|\mathbf{P}_M^{\tilde{\mathbf{A}}^{(k)}} - \theta\|_F^2 \le (\frac{n}{K-1})^2 \tag{A.15}$$

holds with probability one. Plugging Eq. (A.14) and inequality (A.15) into (A.13), we have

$$|V\left(\overset{*}{\mathbf{P}}{}_M^{\tilde{\mathbf{A}}^{(k)}}, \mathbf{P}_0\right) - MSE\left(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0\right)| = O_p(\frac{1}{K}).$$

Since that $V_M(\mathbf{P}_0) = \frac{1}{K}\sum_{k=1}^{K} V\left(\overset{*}{\mathbf{P}}_M^{\tilde{\mathbf{A}}^{(k)}}, \mathbf{P}_0\right)$, we can easily derive

$$|V_M(\mathbf{P}_0) - MSE\left(\mathbf{P}_M^{\mathbf{A}}, \mathbf{P}_0\right)| = O_p(\frac{1}{K}). \tag{A.16}$$

Combining (A.12) and (A.16), we have

$$\text{MMV}(M|\mathbf{A}) - MSE\left(\mathbf{P}_{M_o}^{\mathbf{A}}, \mathbf{P}_0\right) \;=\; O_p(\frac{1}{n} \vee \frac{1}{K}). \tag{A.17}$$

A.2.2 Proof of Theorem 2

We consider two scenarios: (1) Scenario 1: $MSE(\mathbf{P}_{M_o}^{\mathbf{A}}, \mathbf{P}_0) < MSE\left(\mathbf{P}_{M_c}^{\mathbf{A}}, \mathbf{P}_0\right)$, (2) Scenario 2: $MSE\left(\mathbf{P}_{M_o}^{\mathbf{A}}, \mathbf{P}_0\right) \geq MSE\left(\mathbf{P}_{M_c}^{\mathbf{A}}, \mathbf{P}_0\right)$. Under scenario 1, we can directly get $MSE\left(\mathbf{P}_{M_o}^{\mathbf{A}}, \mathbf{P}_0\right) < MSE\left(\mathbf{P}_{M_c}^{\mathbf{A}}, \mathbf{P}_0\right)$, which completes the proof. We then provide proof under Scenario 2. According to the definition of o_p and the results of Theorem 2, for any ϵ, we have

$$\begin{aligned} & \lim_{n\to\infty} \mathbb{P}(|\text{MMV}_{M_o}(\mathbf{A}) - MSE\left(\mathbf{P}_{M_o}^{\mathbf{A}}, \mathbf{P}_0\right)| > \epsilon \max(\frac{1}{n}, \frac{1}{K})) \\ = \; & \lim_{n\to\infty} \mathbb{P}(|\text{MMV}_{M_c}(\mathbf{A}) - MSE\left(\mathbf{P}_{M_c}^{\mathbf{A}}, \mathbf{P}_0\right)| > \epsilon \max(\frac{1}{n}, \frac{1}{K})) \\ = \; & 0. \end{aligned} \tag{A.18}$$

To ease the description, Let MMV_0 denote $\text{MMV}_{M_o}(\mathbf{A})$, MMV_1 denote $\text{MMV}_{M_c}(\mathbf{A})$, L_0 denote $MSE\left(\mathbf{P}^{\mathbf{A}}_{M_o}, \mathbf{P}_0\right)$, L_1 denote $MSE\left(\mathbf{P}^{\mathbf{A}}_{M}, \mathbf{P}_0\right)$, $\Gamma = \max(\frac{1}{n}, \frac{1}{K})$. We have

$$
\begin{aligned}
& \mathbb{P}(L_1 < L_0) \\
= \; & \mathbb{P}(\text{MMV}_1 - \text{MMV}_0 < \{\text{MMV}_1 - L_1\} - \{\text{MMV}_0 - L_0\}) \\
\leq \; & \mathbb{P}(\text{MMV}_1 - \text{MMV}_0 < |\text{MMV}_0 - L_0| + |\text{MMV}_1 - L_1|) \\
\leq \; & \mathbb{P}(\text{MMV}_1 - \text{MMV}_0 < |\text{MMV}_0 - L_0| + |\text{MMV}_1 - L_1| \Big| |\text{MMV}_0 - L_0| > \epsilon\Gamma)\mathbb{P}(|\text{MMV}_0 - L_0| > \epsilon\Gamma) \\
+ \; & \mathbb{P}(\text{MMV}_1 - \text{MMV}_0 < |\text{MMV}_0 - L_0| + |\text{MMV}_1 - L_1| \Big| |\text{MMV}_0 - L_0| \leq \epsilon\Gamma)\mathbb{P}(|\text{MMV}_0 - L_0| \leq \epsilon\Gamma) \\
\leq \; & \mathbb{P}(|\text{MMV}_0 - L_0| > \epsilon\Gamma) + \mathbb{P}(\text{MMV}_1 - \text{MMV}_0 < \epsilon\Gamma + |\text{MMV}_1 - L_1|) \\
\leq \; & \mathbb{P}(|\text{MMV}_0 - L_0| \geq \epsilon\Gamma) + \mathbb{P}(|\text{MMV}_1 - L_1| > \epsilon\Gamma) + \mathbb{P}(\text{MMV}_1 - \text{MMV}_0 < 2\epsilon\Gamma).
\end{aligned}
$$

Under Assumption 3 and the equation (A.18), we have

$$
\begin{aligned}
& \lim_{n\to\infty} \mathbb{P}(L_1 < L_0) \\
\leq \; & \lim_{n\to\infty} \mathbb{P}(|\text{MMV}_0 - L_0| > \epsilon\Gamma) + \lim_{n\to\infty} \mathbb{P}(|\text{MMV}_1 - L_1| > \epsilon\Gamma) + \lim_{n\to\infty} \mathbb{P}(\text{MMV}_1 - \text{MMV}_0 < 2\epsilon\Gamma) \\
= \; & 0,
\end{aligned}
\tag{A.19}
$$

which completes the proof.

A.3 Additional Simulation Results

A.3.1 Sensitivity analysis

In this subsection, we conducted a simulation study to illustrate the impact of λ and θ on the performance of MMV. All results are obtained from 100 replications.

Sensitivity analysis of λ. Figure A.1 shows the effects of varying $\lambda \in \{0.1, \ldots, 0.9\}$ on method selection under $n = 200$ in method selection. As we can see, the method selection accuracy is not sensitive to λ under all graphons. In addition, Figure A.2 show the computational cost of varying $\lambda \in$

$\{0.1, \ldots, 0.9\}$. As we expected, when λ increases, the computational cost significantly increases. Thus, we set $\lambda = 0.1$ in this paper to ease the computational cost.

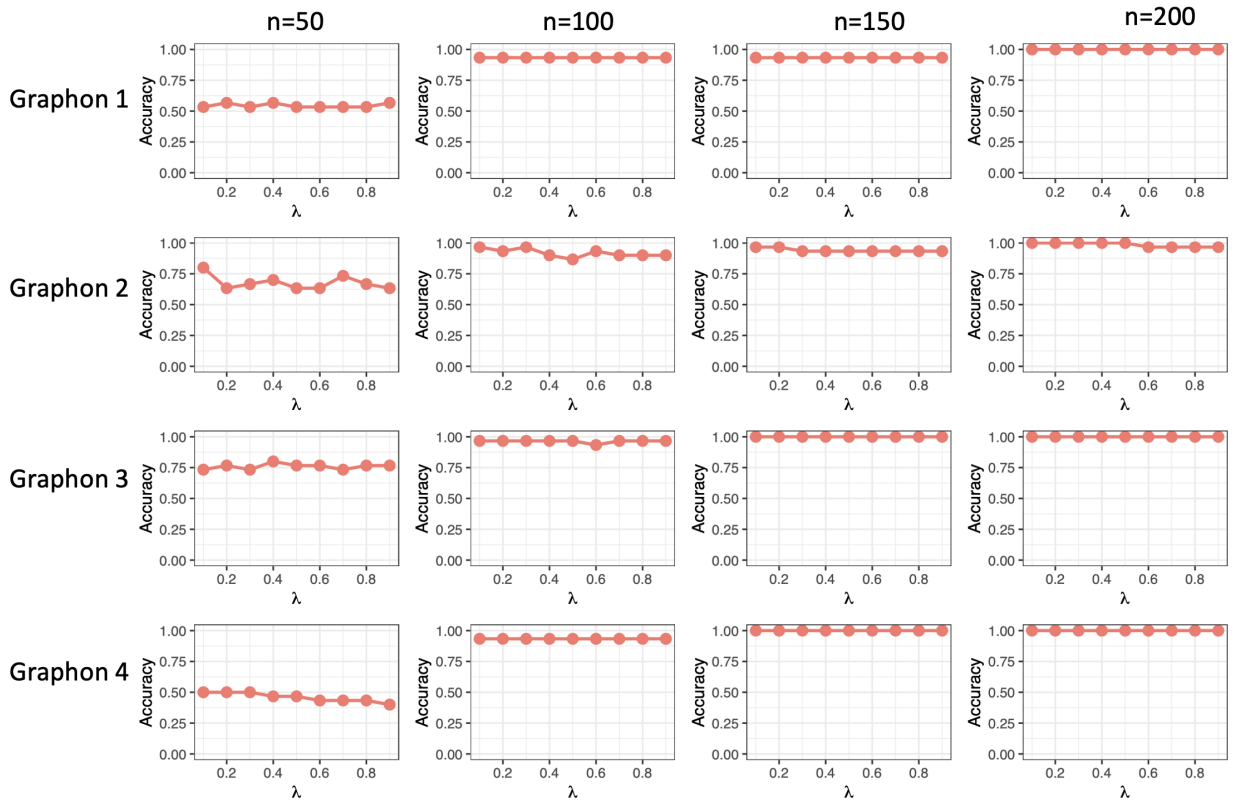


Figure A.1: Method selection accuracy with different λ under $n = 50, 100, 150, 200$ (from left to right) for graphons 1-4 (from top to bottom).

Sensitivity analysis of θ. Recall that in this paper, we propose a data-driven way to decide θ, i.e., setting set θ to the network density. In this subsection, we conduct simulation studies to confirm this proposal. Figure A.3 shows the effects of varying $\theta \in \{0.1, \ldots, 0.9\}$ in method selection. The black dashed line locates the approximated network density, i.e., 0.9, 0.6, 0.3, 0.3 for graphons 1-4, which is observed in Table 1. From Figure A.3, we observe that when n is small, such as 50 and 100, the method selection accuracy is affected by θ, but when n increases to 150, the method selection accuracy is barely affected by θ. In addition, when $n = 50$ or $n = 100$, the accuracy of θ where the black dashed lines locates always achieves a fairly competitive accuracy by successfully avoiding the bad choice of θ values. These observations empirically confirm our proposal.

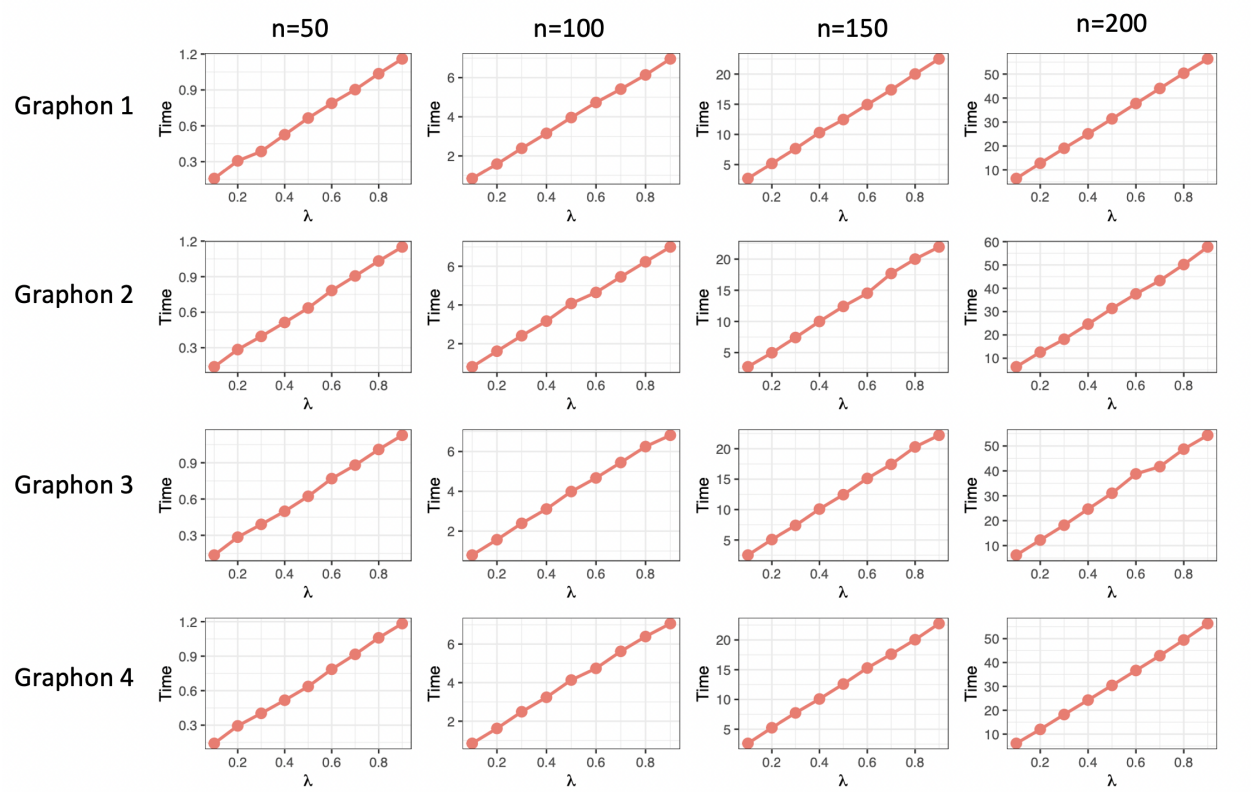


Figure A.2: Method selection computational time with different λ under $n = 50, 100, 150, 200$ (from left to right) for graphons 1-4 (from top to bottom).

A.3.2 Assumption validation

In this subsection, we empirically show that prominent graphon estimators (Airoldi et al., 2013; Chan & Airoldi, 2014; Chatterjee, 2015; Qin et al., 2021; Y. Zhang et al., 2017) satisfy Assumption 2. In particular, we let $D = \frac{\|\mathbf{P}_M^{\mathbf{A}} - \mathbf{P}_M^{\tilde{\mathbf{A}}}\|_F}{\|\tilde{\mathbf{A}} - \mathbf{A}\|_F}$, and conducted extensive experiments to investigate the how D changes as n increases.

In particular, we follow the same simulation settings to generate $\mathbf{A}$ with the number of nodes n varying from 50 to 350. We set $K = 0.1n$. We get estimator $\mathbf{P}_M^{\mathbf{A}}$ and $\mathbf{P}_M^{\tilde{\mathbf{A}}}$ using five prominent graphon estimation methods: (1) Stochastic block model approximation (SBA) algorithm (Airoldi et al., 2013), (2) Sort-and-smooth (SAS) method (Chan & Airoldi, 2014), (3) Universal singular value thresholding (USVT) algorithm (Chatterjee, 2015), (4) Neighborhood smoothing (NS) method (Y. Zhang et al., 2017)

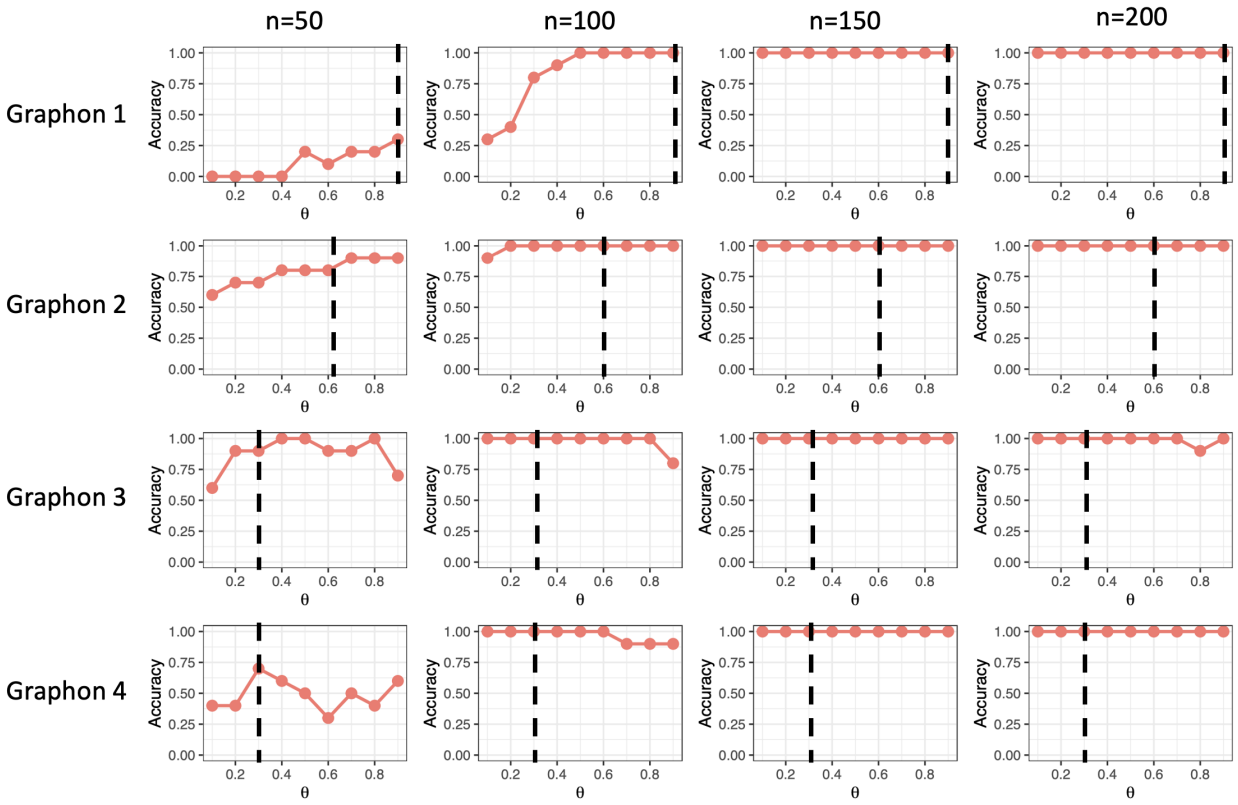


Figure A.3: Method selection accuracy with different θ under $n = 50, 100, 150, 200$ (from left to right) for graphons 1-4 (from top to bottom). The black dashed line locates the approximated network density.

(5) Iterative connecting probability estimation method (ICE) (Qin et al., 2021). Here, we implement the aforementioned methods using the default hyperparameter.

Figure A.4 shows the trend of D with n increases under different graphon settings. All results are averaged over 100 replications. As we can see, for all five graphon estimation methods considered, and under all graphon settings, we observe that D decreases with n, indicating that D is bounded. This observation empirically validates that Assumption 2 is satisfied.

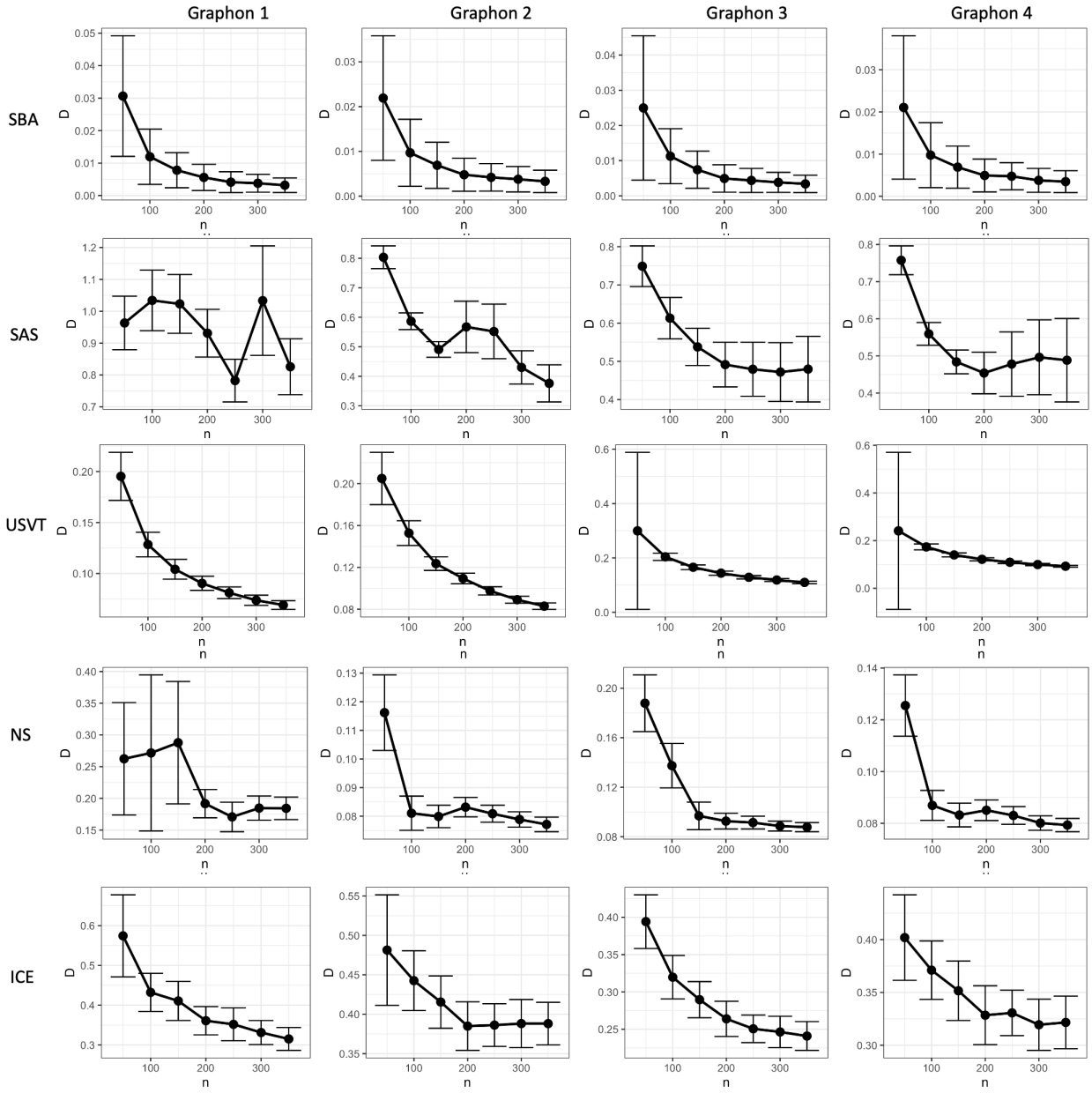


Figure A.4: Averaged D with error bar over 100 replications. The lowest and the highest value of the error bar represent the mean $\pm$ standard deviation. From top to bottom, we show results obtained by applying five different graphon estimation methods, i.e., SBA, SAS, USVT, NS, and ICE.

Appendix B

Appendix for Chapter 3

The appendix shows details of experiments, including the parameter settings of generating the synthetic datasets, a description of real-world datasets, and more experiment results on synthetic and real-world datasets.

B.1 Details of Experiments

We evaluate the performance of our algorithm on both synthetic datasets and real-world datasets using the metrics presented in the main context. We also compare the proposed method with seven benchmark exploration-based graph subsampling methods, including Metropolis Hasting Random Walk Sampler (MHRW), Forest Fire Sampler (FFS), Snowball Sampler, Community Structure Expansion Sampler (CSE), degree-based node sampler (DBN), Random Walk Sampler (RW) and Multi-dimensional Random Walk Sampler (MDRW). We set the hyper-parameters of the seven methods as the default in the package *Little Ball of Fur*. The rejection constraint of the MHRW method is set as 1; the burning probability of FFS is 0.4; the bound on the degree of Snowball Sampler is set as 50. The hyper-parameter used to calculate the Ricci curvature of graphs is set as $\alpha = 0.5$. The times of subsampling r in Algorithm 2 for estimation of the number of communities is set as 3, which is enough to get a stable estimation result. All the experiments are conducted on a workstation with a 40-core NVIDIA Tesla V100 GPU (3.00 GHz).

B.1.1 Experiments on Synthetic Datasets

Parameters Settings

We use stochastic block models (SBM), and degree corrected block models (DCBM), which can assign the community distribution to each node. To create graphs with unbalanced communities, we set the community proportion as $(3/4, 1/10, 1/12, 1/15)$ with 900 nodes. The out-in-ratio controls the ratio of an edge's probability of between-communities and within-community. A higher out-in-ratio represents a noisier graph, i.e., harder to distinguish the community from the graph. We set the probability of an edge within a block as 0.8 and vary the probability of an edge between blocks (pout) for both SBM and DCBM ($\{0.06, 0.08, 0.10, 0.12\}$). The higher the pout value (higher the out-in-ratio) is, the noisier the subsampled graph is. We also change the subsampling proportion ($\{0.1, 0.12, 0.14, 0.16\}$) from low to high for both block models to observe how our method performs compared with others as the settings change. The degree corrected model corrects the node degree using a power-law distribution. More parameters need to be set before generating the graphs following DCBM. We set the average node degree as 40, with is consistent with the setting in Li et al., 2020, in which the assumptions in our algorithm hold. The node degree follows a power-law distribution with the lower bound as 1 and the scaling parameter as 5.

Additional Results of Synthetic Datasets

We use synthetic datasets with hidden community structures to evaluate the performance of our method. We replicate the experiments 30 times under each setting and compare the performance of the error of estimation of the number of communities and the computation time. The average of 30 replications of each setting is recorded in Table B.1 and Table B.2. In these tables, Column **Prop** denotes the subsampling proportion, and column **Pout** denotes the probability of an edge between communities. The rest of the columns are the results of different subsampling methods. The performance of our algorithm is better than that of other methods. As for the results for SBM in Table B.2, the error of estimating the

Table B.1: DCBM: Error of the estimation of the number of communities for different subsampling methods under different settings.

Pout	Prop	ORG-sub	MHRW	CSE	FFS	SnowBall	DBN	RW	MDRW
0.06	0.10	**1.50**	2.83	2.83	2.73	1.53	2.60	2.77	3.00
	0.12	**1.40**	2.77	2.77	2.60	1.90	2.63	2.73	3.00
	0.14	**1.57**	2.53	2.73	2.33	1.80	2.57	2.57	3.00
	0.16	**1.63**	2.50	2.80	2.23	1.80	2.30	2.53	3.00
0.08	0.10	**1.30**	2.97	2.83	2.70	1.53	2.77	2.80	3.00
	0.12	**1.57**	2.77	2.83	2.77	1.80	2.53	2.53	3.00
	0.14	**1.57**	2.90	2.83	2.70	1.80	2.40	2.67	3.00
	0.16	**1.63**	2.77	2.63	2.53	1.80	2.40	2.53	3.00
0.10	0.10	**0.93**	2.83	2.93	2.83	1.37	2.63	2.80	3.00
	0.12	**1.63**	2.90	2.87	2.60	1.83	2.50	2.73	3.00
	0.14	**1.53**	2.80	2.77	2.70	1.93	2.47	2.77	3.00
	0.16	**1.73**	2.83	2.77	2.57	1.70	2.57	2.63	3.00
0.12	0.10	**0.93**	2.93	2.90	2.73	1.6	2.77	2.93	3.00
	0.12	**1.40**	3.00	2.77	2.73	1.60	2.73	2.80	3.00
	0.14	**1.57**	2.93	2.73	2.47	1.80	2.73	2.77	3.00
	0.16	**1.60**	2.90	2.67	2.50	1.87	2.60	2.67	3.00

number of communities decreases as the subsampling proportion increases, and the error increases as the observed graphs are noisier (pout increases). As for the computational time recorded in Table B.3 for SBM, we observe that the estimation time of the full sample is two orders of magnitude larger than the time of subsampling and estimation. As the subsampling proportion increases, the computational time of subsampling methods increases. Though our method's computational time is slightly larger than other subsampling methods, our method has better accuracy.

B.1.2 Experiments on Real-world Datasets

Description of the Datasets

We use five real-world datasets to evaluate the performance of our method as well. These datasets are widely used as benchmark datasets in the study of community detection for graphs. They can be easily found at KONECT, UCI network data repository and Stanford Network Analysis Project Rossi and Ahmed, 2015;

Table B.2: SBM: Error of the estimation of the number of communities for different subsampling methods under different settings.

Pout	Prop	ORG-sub	MHRW	CSE	FFS	SnowBall	DBN	RW	MDRW
	0.10	**0.47**	1.70	1.93	2.00	2.33	2.00	2.00	3.00
0.06	0.12	**0.33**	1.67	1.80	1.93	2.20	2.00	2.00	3.00
	0.14	**0.10**	1.77	1.77	1.93	2.13	2.00	2.00	3.00
	0.16	**0.00**	1.73	1.30	1.97	2.07	2.00	2.00	3.00
	0.10	**0.63**	1.77	1.60	1.87	2.23	2.00	2.00	3.00
0.08	0.12	**0.37**	1.73	1.70	1.83	2.17	2.00	2.00	3.00
	0.14	**0.03**	1.87	1.53	1.80	2.07	2.00	2.00	3.00
	0.16	**0.07**	1.83	1.67	1.83	2.03	2.00	2.00	3.00
	0.10	**0.80**	1.77	1.73	1.93	2.20	2.00	1.97	3.00
0.10	0.12	**0.33**	1.73	1.67	2.00	2.17	2.00	1.93	3.00
	0.14	**0.13**	1.80	1.63	1.90	2.03	2.00	1.93	3.00
	0.16	**0.03**	1.73	1.60	1.97	2.03	2.00	1.93	3.00
	0.10	**0.73**	1.67	1.83	1.87	2.07	2.00	2.00	3.00
0.12	0.12	**0.53**	1.60	1.57	1.93	2.03	2.00	2.00	3.00
	0.14	**0.33**	1.63	1.53	1.93	2.00	2.00	2.00	3.00
	0.16	**0.20**	1.63	1.40	1.97	2.00	2.00	2.00	3.00

Rozemberczki et al., 2020; Sen et al., 2008. All datasets are considered undirected and unweighted graphs, and all self-loop edges and isolated nodes are removed. More details about the datasets are as follow.

Polblogs: The political blogs network dataset was collected by Adamic and Glance in 2005. A network is constructed among all the posts published by liberal or conservative bloggers (two communities). Each node represents one post. An edge connects two nodes if one of them is cited by the other.

Polbooks: This is a network of books about US politics published around the 2004 presidential election and sold by online bookseller *Amazon.com*. All the books are divided into four communities by NI-LPA. Edges between books represent frequent co-purchasing of books by the same buyers.

Facebook: This is an ego-network dataset of the "friend circles" of one anonymous user on Facebook. The network forms friend circles such as family members, high school friends or other friends that are "hand-labeled" by the user.

Table B.3: SBM: Comparison of the computation time (s) of the estimation of the number of communities between using the full dataset and the sampled dataset for different subsampling methods under different settings (seconds).

Pout	Full	Prop	ORG-sub	MHRW	CSE	FFS	SnowBall	DBN	RW	MDRW
		0.10	1.11	0.42	3.27	0.43	0.44	0.29	0.37	0.28
0.06	**22.52**	0.12	1.21	0.48	3.76	0.49	0.52	0.49	0.54	0.37
		0.14	1.41	0.54	4.41	0.53	0.62	0.21	0.33	0.23
		0.16	1.63	0.66	5.12	0.66	0.67	0.42	0.44	0.38
		0.10	1.08	0.42	3.21	0.42	0.43	0.32	0.41	0.26
0.08	**23.67**	0.12	1.27	0.49	3.87	0.50	0.53	0.25	0.24	0.20
		0.14	1.47	0.57	4.50	0.54	0.61	0.25	0.26	0.25
		0.16	1.71	0.65	5.23	0.66	0.68	0.53	0.67	0.50
		0.10	1.09	0.41	3.25	0.42	0.43	0.38	0.42	0.32
0.10	**15.67**	0.12	1.30	0.49	3.91	0.50	0.53	0.27	0.28	0.17
		0.14	1.52	0.56	4.57	0.54	0.61	0.26	0.34	0.26
		0.16	1.78	0.65	5.28	0.66	0.68	0.57	0.68	0.52
		0.10	1.13	0.42	3.28	0.42	0.45	0.18	0.24	0.22
0.12	**19.65**	0.12	1.38	0.48	4.11	0.50	0.53	0.24	0.32	0.20
		0.14	1.61	0.56	4.71	0.54	0.61	0.45	0.56	0.43
		0.16	1.89	0.65	5.39	0.68	0.68	0.41	0.49	0.43

Cora: The Cora dataset describes the citation relationship among scientific publications classified into seven classes. After preprocessing, there remain 2,485 nodes and 5,069 links.

PubMed: The PubMed dataset describes the citation relationship among scientific publications classified into seven classes. After preprocessing, there remain 19,717 nodes and 44,338 links.

Table B.4 presents the summary statistics for those real networks we used in this paper. As we can see, these datasets cover different levels of network size, density, number fo communities, clustering coefficient (CC), and imbalance. The degree distributions of all the datasets are presented in Figure B.1.

Computational Time

The comparisons of errors in estimating the number of communities are presented in the main context. The computation time is also obtained in B.5. Similar to the synthetic datasets, the estimation time of

Table B.4: Key features of the real-world datasets. Node and Edge represent the number of nodes, edges, and communities of the graph, respectively. Density is the edge density, calculated by the ratio of the number of edges in the actual and complete graphs. CC is the clustering coefficient of the whole graph. IM is the imbalance level.

Dataset	Node	Edge	Density	Community	CC	IM
Polblogs	1224	16718	0.0112	2	0.2260	0.72
Polbooks	105	441	0.0400	3	0.3484	0.89
Facebook	725	13030	0.0248	11	0.4576	0.71
Cora	2485	5069	0.0008	7	0.0900	0.94
PubMed	19717	44338	0.0001	3	0.0537	0.96

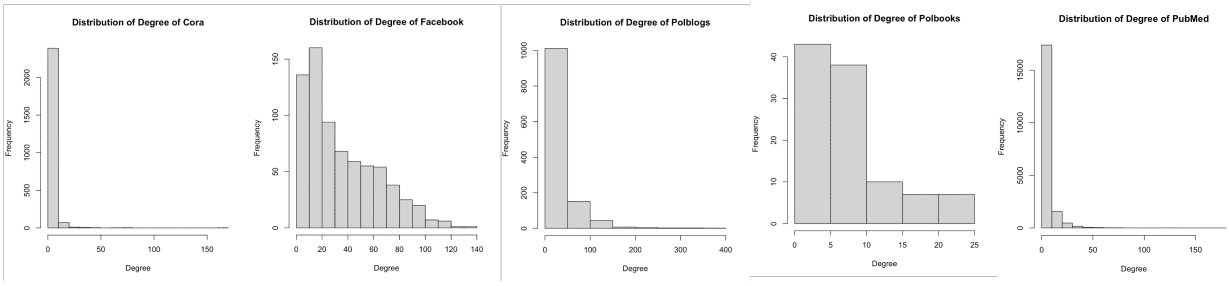


Figure B.1: The degree distribution of the five datasets.

Table B.5: Comparison of the computation time (seconds) of the estimation of the number of communities between using the full dataset and the sampled dataset for each dataset and subsample method with different subsample size.

Dataset	Full	Ricci	Prop	ORG-sub	MHRW	CSE	FFS	SnowBall	DBN	RW	MDRW
			10%	0.10	0.05	0.07	0.08	0.10	0.05	0.07	0.04
Polbooks	**1.88**	0.29	20%	0.13	0.10	0.11	0.11	0.11	0.06	0.11	0.06
			30%	0.14	0.11	0.12	0.12	0.12	0.10	0.11	0.07
			10%	0.65	0.50	0.74	0.51	0.53	0.47	0.49	0.47
Facebook	**29.14**	0.71	20%	1.01	0.75	1.41	0.78	0.83	0.71	0.71	0.71
			30%	1.96	1.56	2.83	1.70	1.69	1.48	1.53	1.49
			5%	0.34	0.23	0.49	0.24	0.22	0.22	0.22	0.29
Cora	**191.06**	0.79	10%	0.79	0.66	1.54	0.65	0.56	0.49	0.60	0.73
			20%	2.30	1.88	4.83	1.90	1.79	1.65	1.93	2.21
			5%	0.23	0.12	0.13	0.23	0.13	0.11	0.22	0.12
Polblogs	**48.6**	1.49	10%	0.56	0.27	0.30	0.67	0.29	0.35	0.42	0.31
			20%	1.18	0.49	0.47	1.01	0.48	0.50	0.67	0.44
			0.5%	1.03	0.64	1.23	0.57	0.53	0.79	0.59	0.61
PubMed	**NA**	41.59	2%	4.42	3.93	11.84	4.13	3.50	2.63	3.58	3.43
			5%	8.92	8.61	43.36	8.94	8.24	6.62	8.83	7.85

the full sample is two orders of magnitude larger than the time of subsampling (including the time for computing OR curvature for each edge in the graph) and estimation.

The result also shows that the computation time for estimating the number of communities using the subsampling method is much shorter than that using the full dataset. The complexity of all methods is influenced by the size of the graph, especially by the number of nodes. It is no wonder that the computation time of all methods in estimating the number of communities for dataset *Cora* is much longer than other datasets.

B.1.3 Additional empirical results

In this section, we provide additional empirical results. In real-world networks, the ground truth is not available. We used some data sets that are widely used for community detection tasks, where the communities are labeled with domain knowledge. The manually labeled community structures correspond to different underlying aspects of the nodes. For example, in the Cora citation network (where a node represents a paper), the community label corresponds to the paper's subject (e.g., neural network). We agree that the manually labeled number of communities might not be the ground truth, which is a limitation for real networks.

To overcome this limitation, we also employ another evaluation metric to assess the performance of our methods. In particular, we do not compare our estimation with the manually labeled number of communities. We compare our estimation with the estimation obtained from the full data. These comparison results can be used to evaluate whether the subsampled subgraph can be a good surrogate to carry out computations of interest for the full data. Table B.6 shows the estimation difference between the subsampled and full graphs for four datasets, i.e., Polbooks, Facebook, Cora, and Polblogs. Here we do not show the results for PubMed, since the estimation results of PubMed full graph are not available due to prohibitive computation. As we can see, using this new metric, our method ORG-sub still outperforms other subsampling methods.

Table B.6: The estimation difference between sampling and full.

Dataset	Prop.	ORG-sub	MHRW	CSE	FFS	SnowBall	DBN	RW	MDRW
Polbooks	10%	**0.08**	0.66	0.39	2.79	1.08	0.29	0.20	1.00
	20%	0.29	0.46	0.34	0.59	0.37	0.57	**0.28**	0.61
	30%	**0.13**	0.75	0.34	2.86	1.11	1.80	1.70	0.99
Facebook	10%	**1.36**	2.90	3.90	2.96	4.66	3.86	1.67	6.00
	20%	**1.31**	2.84	3.73	2.24	3.09	2.87	1.44	5.94
	30%	**1.20**	2.91	3.70	2.11	3.23	2.71	1.42	5.85
Cora	5%	**0.27**	0.87	0.60	2.37	1.13	0.63	2.03	0.83
	10%	0.80	3.90	0.47	0.80	**0.60**	1.60	1.13	1.33
	20%	0.27	0.90	2.00	**0.03**	1.40	1.77	1.27	2.20
Polblogs	5%	**0.00**	1.87	0.90	2.00	0.43	1.33	1.03	0.30
	10%	**0.00**	0.40	0.33	0.20	0.03	0.03	0.07	0.87
	20%	**0.00**	1.87	0.90	2.00	0.43	1.33	1.03	0.30

B.2 Discussion on Edge Density

Indeed, the performance of our method depends on the edge density from both theory and empirical results. On one hand, Corollary 4.1.2 basically assumes that $\rho > \frac{log(\tilde{n})}{\tilde{n}}$, where $\tilde{n}$ is the number of sampled nodes, and ρ is the edge density. When $\tilde{n} = 100$, we require $\rho > 0.046$, and when $\tilde{n} = 1000$, we only require $\rho > 0.007$. As the subsample size becomes larger, the constraint imposed on the edge density becomes weaker. On the other hand, we include additional simulation studies to show the empirical performance of our method with different network densities. Particularly, we use the same SBM setting in Section 5.1 to generate synthetic data, except that we let $p_{in} = 0.8\lambda$ and $p_{out} = 0.1\lambda$, where $\lambda \in \{0.2, 0.4, \ldots, 1\}$ controls the edge density level. Here, we fix the sampling proportion as 10%, that is, we sample 90 nodes. Table B.7 shows the edge density and average estimation error under different λ. As we can see, an increasing edge density comes with a lower error.

Table B.7: SBM:Estimation error of the number of communities with different edge density under the SBM model.

λ	0.2	0.4	0.6	0.8	1.0
Density	0.1018	0.2034	0.3053	0.4070	0.5085
Error	2.7	2.1	1.8	1.8	1.5

B.3 Related Work on Graph Subsampling

Existing graph subsampling techniques can be categorized into three main groups: node-based, edge-based, and exploration-based.

Node-based sampling. Random Node (RN) sampling (Leskovec & Faloutsos, 2006) and Degree-based Node sample (DBN) (Leskovec & Faloutsos, 2006) sampling are two most common node-base sampling methods. RN selects a set of nodes uniformly at random from the graph, while DBN selects a node with a probability that is proportional to its degree. DBN has been shown to favor high-degree nodes.

Edge-based sampling. Random Edge (RE) sampling (Leskovec & Faloutsos, 2006) generates an induced subgraph by selecting edges uniformly at random. Some variants of RE have been proposed, such as Random Node-Edge (RNE) sampling (Rafiei, 2005) that randomly selects a node and then randomly chooses an adjacent edge. Previous studies (Leskovec & Faloutsos, 2006) have demonstrated that neither RE nor RNE preserves community structures because the resulting sampled graphs are often sparsely connected. Meanwhile, both RE and REN slightly favor high-degree nodes because the probability of selecting a node increases with its degree.

Exploration-based sampling. SnowBall sampling (Goodman, 1961) and Forest Fire sampling (FFS) (Leskovec et al., 2005) are two basic exploration-based sampling methods, selecting a fixed fraction of neighbors visited at each iteration. Random walk (RW) (Lovász, 1996) is also a popular exploration-based graph sampling method, which selects the next node at random from the neighbors of the currently selected node. One of the major limitations of RW is that RW is inherently biased towards visiting high-degree nodes. To overcome the drawback of RW, researchers proposed Metropolis Hasting Random Walk Sampler (MHRW) (Hübler et al., 2008) and Multi-dimensional Random Walk Sampler (MDRW) (Ribeiro & Towsley, 2010). Most of the aforementioned literature aims at getting a subgraph that preserves some network summary statistics, such as degree distribution. Only a few works aim at obtaining a subgraph that preserves the community information. (Maiya & Berger-Wolf, 2010) proposed a local greedy

search-based community structure expansion sampling (CSE) method to optimize the preservation of community structures.

Despite many successful applications of existing graph sampling methods. They all have various limitations. In particular, node-based and edge-based sampling methods sample nodes or edges independently, ignoring the neighborhoods of seed nodes. They might obtain a disconnected subgraph from a connected graph (Wu et al., 2016). For RW and FF, it has been shown that they could get trapped inside the communities and leave other communities out of the sample entirely (Wu et al., 2016). While the MHRW algorithm ensures that the subgraph preserves degree distribution, the performance of this algorithm is dependent on its sample acceptance ratio. It has been shown that the acceptance ratio of MHRW is typically very low in real-world networks **empty citation** Therefore, MHRW generally suffers from the sample rejection problem, which clearly degrades the performance of MHRW. For CSE, the selection procedure of the next node is not based on a mathematical framework and one cannot compute the probability of visiting sampled nodes in CSE (Salehi et al., 2012).

B.4 Empirical Investigation of Future Work

Theoretically, the OR curvature has been proven to be related to some network summaries concerning communities, such as the eigenvalues of the graph Laplacian Bauer et al., 2011 and the clustering coefficient (CC) Jost and Liu, 2014.

Empirically, we calculated the CC value of the full graph and that of the subsampled graph for two real datasets: Facebook and Cora. We set the subsampling proportion to 10%. We replicated the sampling procedure 100 times and thus obtained 100 subsamples. Using these subsamples, we obtained the 95% confidence interval (CI) by calculating mean $\pm$ 1.96*sd (standard deviation).

For Facebook data, the CI of the CC value is [0.420, 0.537], and the CC value obtained from the full graph is 0.476. For Cora data, the CI is [0.234, 0.290], and the CC value of the full graph is 0.238. As we can see, the 95% CI covers the CC value of the full graph for both real data sets. These observations

suggest that our subsampling algorithm could preserve CC to some extent. The consistency theory on CC using our method is under investigation, and the results will be reported in future publications.

Bibliography

Adler, E. (2005). Epistemic communities. *International encyclopedia of governance*, *1*, 115–122.

Adler, P. (2005). The changing nature of work: Implications for occupational analysis. *Research in the Sociology of Work*, *15*, 17–43.

Airoldi, E. M., Costa, T. B., & Chan, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. *Advances in Neural Information Processing Systems*, *26*, 692–700.

Alzahrani, T., & Horadam, K. J. (2016). Community detection in bipartite networks: Algorithms and case studies. In *Complex systems and networks* (pp. 25–50). Springer.

Avant, D., & Westerwinter, O. (2016). *The new power politics: Networks and transnational security governance*. Oxford University Press.

Barabási, A.-L. (2016). *Network science*. Cambridge University Press.

Bauer, F., Jost, J., & Liu, S. (2011). Ollivier-ricci curvature and the spectrum of the normalized graph laplace operator. *arXiv preprint arXiv:1105.3803*.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Bob, C. (2005). *The marketing of rebellion: Insurgents, media, and international activism*. Cambridge University Press.

Boli, J., & Thomas, G. M. (1999). *Constructing world culture: International nongovernmental organizations since 1875*. Stanford University Press.

Brechenmacher, S. (2017). *Civil society under assault: Repression and responses in russia, egypt, and ethiopia* (Vol. 18). Carnegie Endowment for International Peace Washington, DC.

Brysk, A. (1993a). From above and below: Social movements, the international system, and human rights in argentina. *Comparative Political Studies*, *26*(3), 259–285.

Brysk, A. (1993b). The politics of human rights in argentina: Protest, change, and democratization. *Stanford University Press*.

Bueger, C., & Gadinger, F. (2015). The play of international practice. *International Studies Quarterly*, *59*(3), 449–460. https://doi.org/10.1111/isqu.12189

Burt, R. S. (2003). The social structure of competition. *Networks in the knowledge economy*, *13*, 57–91.

Bush, S. S. (2015). *The taming of democracy assistance*. Cambridge University Press.

Butts, C. T., et al. (2008). Social network analysis with sna. *Journal of statistical software*, *24*(6), 1–51.

Caniglia, B. (2001). Informal alliances vs. institutional ties: The effects of elite alliances on environmental tsmo networks. *Mobilization: An International Quarterly*, *6*(1), 37–54.

Carpenter, C. (2014). *"lost" causes: Agenda vetting in global issue networks and the shaping of human security*. Cornell University Press. http://www.jstor.org/stable/10.7591/j.ctt5hhor5

Chan, S., & Airoldi, E. (2014). A consistent histogram estimator for exchangeable graph models. *International Conference on Machine Learning*, 208–216.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, *43*(1), 177–214.

Cheng, H.-M., Ning, Y.-Z., Yin, Z., Yan, C., Liu, X., & Zhang, Z.-Y. (2018). Community detection in complex networks using link prediction. *Modern Physics Letters B*, *32*(01), 1850004.

Cooley, A., & Ron, J. (2002). The ngo scramble: Organizational insecurity and the political economy of transnational action. *International security*, *27*(1), 5–39.

Deloffre, M. (2016). Towards an organizational sociology of international organizations. *International Political Sociology*, *10*(1), 16–33.

DeMars, W. E. (2005a). Ngos and transnational networks: Wild cards in world politics. *Annals of the American Academy of Political and Social Science*, *598*(1), 200–218. https://doi.org/10.1177/0002716204270336

DeMars, W. E. (2005b). *Ngos and transnational networks: Wild cards in world politics*. Pluto Press London.

Djelic, M., & Quack, S. (2010). Transnational communities: Shaping global economic governance. *Cambridge University Press*.

Djelic, M., & Quack, S. (2011). Transnational communities and the evolution of global governance. *Journal of Business Ethics*, *100*(S1), 85–101.

Djelic, M.-L., & Quack, S. (2010a). Globalization and institutions: Redefining the rules of the economic game. *International Studies Quarterly*, *54*, 375–397.

Djelic, M.-L., & Quack, S. (2010b). *Transnational governance: Institutional dynamics of regulation*. Cambridge University Press.

Fan, J., Fan, Y., Han, X., & Lv, J. (2019). Simple: Statistical inference on membership profiles in large networks. *arXiv preprint arXiv:1910.01734*.

Fernandez, R. M., & Gould, R. V. (1994). A dilemma of state power: Brokerage and influence in the national health policy domain. *American journal of Sociology*, *99*(6), 1455–1491.

Fowler, A. (2000). Ngo futures: Beyond aid: Ngdo values and the fourth position. *Third World Quarterly*, *21*(4), 589–603. http://www.jstor.org/stable/3993366

Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, *30*(4), 1141–1144.

Goodman, L. A. (1961). Snowball sampling. *The annals of mathematical statistics*, 148–170.

Gould, R. V., & Fernandez, R. M. (1989a). Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological Methodology*, *19*, 89–126.

Gould, R. V., & Fernandez, R. M. (1989b). Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological methodology*, 89–126.

Hadden, J., & Bush, S. (2020). Ngo cooperation in global governance. *Annual Review of Political Science*, *23*, 129–147.

Hadden, J., & Jasny, L. (2019). The power of peers: How transnational advocacy networks shape ngo strategies on climate change. *British Journal of Political Science*, *49*(2), 637–659.

Hafner-Burton, E. M., Kahler, M., & Montgomery, A. H. (2009). Network analysis for international relations. *International organization*, *63*(3), 559–592.

Henriksen, L. F., & Seabrooke, L. (2016). Transnational organizing: Issue professionals in environmental sustainability networks. *Organization*, *23*(5), 722–741.

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, *5*(2), 109–137.

Hübler, C., Kriegel, H.-P., Borgwardt, K., & Ghahramani, Z. (2008). Metropolis algorithms for representative subgraph sampling. *2008 Eighth IEEE International Conference on Data Mining*, 283–292.

Hutchison, E. (2016). The role of issue entrepreneurs in networked governance: Some insights from the case of aviation security. *Journal of European Public Policy*, *23*(9), 1296–1313.

Jackson, S. (2020). Towards transformative solidarity: Reflections from amnesty international's global transition programme. *Emory Int'l L. Rev.*, *34*, 705.

Jordan, L., & Van Tuijl, P. (2000). Political responsibility in transnational ngo advocacy. *World development*, *28*(12), 2051–2065.

Jost, J., & Liu, S. (2014). Ollivier's ricci curvature, local clustering and curvature-dimension inequalities on graphs. *Discrete & Computational Geometry*, *51*(2), 300–322.

Kahler, M. (2011). *Networked politics: Agency, power, and governance*. Cornell University Press.

Kassal, I., Jordan, S. P., Love, P. J., Mohseni, M., & Aspuru-Guzik, A. (2008). Polynomial-time quantum algorithm for the simulation of chemical dynamics. *Proceedings of the National Academy of Sciences*, *105*(48), 18681–18686.

Keck, M., & Sikkink, K. (1998). Activists beyond borders: Advocacy networks in international politics. *Cornell University Press*.

Keck, M. E., & Sikkink, K. (1998a). *Activists beyond borders: Advocacy networks in international politics*. Cornell University Press. http://www.jstor.org/stable/10.7591/j.ctt5hh13f

Keck, M. E., & Sikkink, K. (1998b). Activists beyond borders: Advocacy networks in international politics. *Cornell University Press*.

Kolaczyk, E. D., Lin, L., Rosenberg, S., Walters, J., & Xu, J. (2020). Averages of unlabeled networks: Geometric characterization and asymptotic behavior. *The Annals of Statistics*, *48*(1), 514–538.

Koliba, C., & Gajda, R. (2009). Building capacity for public service innovation: The role of communities of practice. *Public Administration Review*, *69*(6), 1075–1086.

Krishnamurthy, V., Faloutsos, M., Chrobak, M., Lao, L., Cui, J.-H., & Percus, A. G. (2005). Reducing large internet topologies for faster simulations. *International Conference on Research in Networking*, 328–341.

Lave, J., & Wenger, E. (1991). Situated learning: Legitimate peripheral participation. *Cambridge University Press*.

Leskovec, J., & Faloutsos, C. (2006). Sampling from large graphs. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 631–636.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 177–187.

Li, T., Levina, E., & Zhu, J. (2020). Network cross-validation by edge sampling. *Biometrika*, *107*(2), 257–276.

Liu, X., Cheng, H.-M., & Zhang, Z.-Y. (2019). Evaluation of community detection methods. *IEEE Transactions on Knowledge and Data Engineering*, *32*(9), 1736–1746.

Lovász, L. (1996). Random walks on graphs: A survey in combinatorics. *Bolyai Society Mathematical Studies*, 353–397.

Lovász, L., & Szegedy, B. (2006). Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, *96*(6), 933–957.

Maiya, A. S., & Berger-Wolf, T. Y. (2010). Sampling community structure. *Proceedings of the 19th international conference on World wide web*, 701–710.

Mall, R., Langone, R., & Suykens, J. A. (2013). Furs: Fast and unique representative subset selection retaining large-scale community structure. *Social Network Analysis and Mining*, *3*(4), 1075–1095.

Maoz, Z. (2017). Network science and international relations. In *Oxford research encyclopedia of politics*.

Moloo, R. (2011). The quest for legitimacy in the united nations: A role for ngos? *UCLA Journal of International Law and Foreign Affairs*, 1–40.

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., & Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *science*, *328*(5980), 876–878.

Murdie, A. (2014). The ties that bind: A network analysis of human rights international nongovernmental organizations. *British Journal of Political Science*, *44*(1), 1–27.

Murdie, A., & Davis, D. R. (2012). Looking in the mirror: Comparing ingo networks across issue areas. *The Review of International Organizations*, *7*(2), 177–202.

Murdie, A., & Polizzi, M. (2017). Human rights and transnational advocacy networks. In *The oxford handbook of political networks*.

Murdie, A. M. (2014a). Helping hands or grabbing hands? determinants of ngo access to intergovernmental organizations. *International Studies Quarterly*, *58*(1), 1–11. https://doi.org/10.1111/isqu.12060

Murdie, A. M. (2014b). International non-governmental organizations and the creation of norms against torture. *International Studies Quarterly*, *58*(3), 555–567.

Nelson, P. J. (1997). Conflict, legitimacy, and effectiveness: Who speaks for whom in transnational ngo networks lobbying the world bank? *Nonprofit and Voluntary Sector Quarterly*, *26*(4), 421–441. https://doi.org/10.1177/0899764097264003

Newman, M. (2018). *Networks*. Oxford University Press.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, *103*(23), 8577–8582.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113.

Okumu, W. (2019). North–south power asymmetry in international development ngos' partnership practices. *Third World Quarterly*, *40*(10), 1744–1762.

Pallas, C. L., & Urpelainen, J. (2013). Mission and interests: The strategic formation and function of north-south ngo campaigns. *Global Governance*, *19*(3), 401–423. http://www.jstor.org/stable/24526201

Qin, Y., Yu, L., & Li, Y. (2021). Iterative connecting probability estimation for networks. *Advances in Neural Information Processing Systems*, *34*.

Rafiei, D. (2005). Effectively visualizing large networks through sampling. *VIS 05. IEEE Visualization, 2005.*, 375–382.

Ribeiro, B., & Towsley, D. (2010). Estimating and sampling graphs with multidimensional random walks. *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 390–403.

Risse-Kappen, T., Risse, T., Ropp, S. C., Sikkink, K., et al. (1999). *The power of human rights: International norms and domestic change*. Cambridge University Press.

Rohe, K., Chatterjee, S., & Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, *39*(4), 1878–1915.

Rossi, R. A., & Ahmed, N. K. (2015). The network data repository with interactive graph analytics and visualization. *AAAI*. https://networkrepository.com

Rozemberczki, B., Kiss, O., & Sarkar, R. (2020). Karate Club: An API Oriented Open-source Python Framework for Unsupervised Learning on Graphs. *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, 3125–3132.

Salehi, M., Rabiee, H. R., & Rajabi, A. (2012). Sampling from complex networks with high community structures. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *22*(2), 023126.

Schneiker, A., & Joachim, J. (2018). From epistemic communities to epistemic governance: Globalization, power and democratic legitimacy. *Review of International Political Economy*, *25*(3), 360–378.

Sen, P., Namata, G. M., Bilgic, M., Getoor, L., Gallagher, B., & Eliassi-Rad, T. (2008). Collective classification in network data. *AI Magazine*, *29*(3), 93–106.

Shumate, M., & Dewitt, L. (2008). The north/south divide in ngo hyperlink networks. *Journal of Computer-Mediated Communication*, *13*(2), 405–428.

Shumate, M., & Dewitt, P. (2008). Building advocacy coalitions in support of global public goods: A case study of the global environment facility. *Policy Studies Journal*, *36*(4), 583–606.

Snyder, J. (2020). From social movement to moral majority: How ngos fuel backlash against international criminal tribunals. *International Organization*, *74*(1), 91–123. https://doi.org/10.1017/S0020818319000336

Stovel, K., & Shaw, L. (2012). Brokerage. *Annual review of sociology*, *38*, 139–158.

Stroup, S. S., & Wong, W. H. (2017). *The authority trap*. Cornell University Press.

Sullivan, B. N., & Stewart, D. (2017). Do connections always help? network brokerage's negative impact on the emergence of status. In *Emergence*. Emerald Publishing Limited.

Tallberg, J., Dellmuth, L. M., Agné, H., & Duit, A. (2018). Ngo influence in international organizations: Information, access and exchange. *British journal of political science*, *48*(1), 213–238.

Tallberg, J., Sommerer, T., & Squatrito, T. (2018). The opening up of international organizations: Transnational access in global governance. *Cambridge University Press*.

Terman, R. (2019). Rewarding resistance: Theorizing defiance to international shaming. *Manuscript, University of Chicago*.

Townsend, J., & Townsend, A. (2004). Accountability, motivation and practice: Ngos north and south. *Social & Cultural Geography*, *5*(2), 271–284. https://doi.org/10.1080/1464936041000169025 9

Tsingou, E. (2015). A knowledge-based approach to transnational advocacy networks. *European Journal of International Relations*, *21*(4), 844–869.

Wang, T., Chen, Y., Zhang, Z., Xu, T., Jin, L., Hui, P., Deng, B., & Li, X. (2011). Understanding graph sampling algorithms for social network analysis. *2011 31st international conference on distributed computing systems workshops*, 123–128.

Wasserman, S., Faust, K., et al. (1994). Social network analysis: Methods and applications.

Wenger, E. (2006). *Communities of practice: A brief introduction*. National College for School Leadership.

Wong, W. H. (2012). *Internal affairs*. Cornell University Press.

Wu, Y., Cao, N., Archambault, D., Shen, Q., Qu, H., & Cui, W. (2016). Evaluation of graph sampling: A visualization perspective. *IEEE transactions on visualization and computer graphics*, *23*(1), 401–410.

You, T., Cheng, H.-M., Ning, Y.-Z., Shia, B.-C., & Zhang, Z.-Y. (2016). Community detection in complex networks using density-based clustering algorithm and manifold learning. *Physica A: Statistical Mechanics and its Applications*, *464*, 221–230.

Zeng, H., Zhou, H., Srivastava, A., Kannan, R., & Prasanna, V. (2019). Graphsaint: Graph sampling based inductive learning method. *International Conference on Learning Representations*.

Zhang, J., & Cao, J. (2017). Finding common modules in a time-varying network with application to the drosophila melanogaster gene regulation network. *Journal of the American Statistical Association*, *112*(519), 994–1008.

Zhang, Y., Levina, E., & Zhu, J. (2017). Estimating network edge probabilities by neighbourhood smoothing. *Biometrika*, *104*(4), 771–783.