

2021

# Uncertainty quantification in noisy networks

---

<https://hdl.handle.net/2144/43181>

*"Downloaded from OpenBU. Boston University's institutional repository."*

BOSTON UNIVERSITY  
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

**UNCERTAINTY QUANTIFICATION IN NOISY  
NETWORKS**

by

**WENRUI LI**

B.S., Shandong University, 2015  
M.S., University of Washington, 2017

Submitted in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

2021

© 2021 by  
WENRUI LI  
All rights reserved

Approved by

First Reader

---

Eric D. Kolaczyk, Ph.D.  
Professor of Statistics

Second Reader

---

Daniel L. Sussman, Ph.D.  
Assistant Professor of Statistics

Third Reader

---

Laura F. White, Ph.D.  
Associate Professor of Biostatistics

## Dedication

This dissertation is dedicated to my family.

## Acknowledgments

First of all, I would like to express my sincere gratitude to my Ph.D. advisors, Eric Kolaczyk, Daniel Sussman and Laura White, for the continuous support of my study and related research, for their patience, motivation, and immense knowledge. My deepest appreciation goes to my primary advisor, Eric, for providing me opportunities to various types of projects, for his relentless encouragement and trust on me even when I question my abilities, for being so supportive during my job hunt, and for caring about my well-being and offering help. I'm extremely grateful to Dan for his insight on technical details and constructive criticism. Many thanks to Laura for guiding me in the field of epidemiology and always being kind.

Furthermore, I would like to extend my thanks to my collaborators. Thanks, Katia Bulekova and Brian Gregor, for their assistance in coding. I would also like to thank Julio Castrillon, Mark Kon and Snezana Milanovic, for guiding me through the projects during the first year of my PhD, and for being supportive and providing suggestions when I was a first-time instructor for a summer course. Special thanks to Julio for sharing his life experiences with me and chatting with me like a friend. Thanks also to Subhabrata Sen for his guidance on the theoretical work. I admire his enthusiasm and attitude to research. I also had great pleasure of working with Nathaniel Josephs, and I'm impressed by his diligence.

In addition, I would like to thank the faculty, the department staff, and my fellow PhDs in the Department of Mathematics and Statistics at Boston University for their great help and assistance during my time here. I wish to sincerely thank Judith Lok for amazing lectures and discussions on causal inference, and for serving as the chair of my dissertation committee. Thanks also to my officemate, Ying Zhang, for her companionship, empathy and encouragement.

Last but not the least, I can not quantify my appreciation for my parents for their endless love and support in all my life's pursuits. I would not have made it this far without them. I would also like to thank my friends and other family relatives for their help and encouragement.

# UNCERTAINTY QUANTIFICATION IN NOISY NETWORKS

WENRUI LI

Boston University, Graduate School of Arts and Sciences, 2021

Major Professor: Eric D. Kolaczyk, Ph.D.  
Professor of Statistics

## ABSTRACT

In recent years there has been an explosion of network data from seemingly all corners of science – from computer science to engineering, from biology to physics, and from finance to sociology. We face analogues of many of the same fundamental types of problems encountered in a ‘Statistics 101’ course when analyzing network data. Despite roughly 20 years of research in the area, one of the fundamental capabilities that we still lack is quantifying uncertainty through propagation of network error. We conduct basic research laying statistical foundations for uncertainty quantification of this type, within a handful of key paradigms, focusing on problems ranging from epidemics to experiments on networks, when at least a few network replicates are available. Specifically, we study causal inference on noisy networks, and estimation of epidemic reproduction numbers in network-based and non-network-based settings. Ultimately, our work will bring critical insight into how ‘noise’ at the level of observed network connectivity impacts critical inferences and decisions derived from data in complex network systems.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Estimation of the Epidemic Branching Factor in Noisy Contact Networks</b>	<b>3</b>
2.1	Introduction . . . . .	3
2.2	Background . . . . .	7
2.2.1	Noise model . . . . .	7
2.2.2	The branching factor in network-based epidemic models . . . . .	9
2.3	Bias and variance of the observed branching factor . . . . .	10
2.3.1	Arbitrary network topology . . . . .	10
2.3.2	Specific network topology . . . . .	11
2.3.3	Simulation study . . . . .	13
2.4	Estimator for the true branching factor . . . . .	13
2.5	Numerical illustration . . . . .	17
2.5.1	Simulations . . . . .	18
2.5.2	Application . . . . .	20
<b>3</b>	<b>Causal Inference under Network Interference with Noise</b>	<b>23</b>
3.1	Introduction . . . . .	23
3.1.1	Problem setup . . . . .	24
3.1.2	Related literature . . . . .	28
3.1.3	Our contributions . . . . .	30
3.2	Impact of ignoring network noise . . . . .	30

3.2.1	Network settings and assumptions . . . . .	31
3.2.2	Biases of standard estimators in noisy networks . . . . .	33
3.2.3	Variances of standard estimators in noisy networks . . . . .	34
3.3	Accounting for network noise . . . . .	35
3.3.1	Method-of-moments estimators . . . . .	35
3.3.2	Asymptotic unbiasedness and consistency . . . . .	38
3.4	Numerical illustration: British secondary school contact networks . . . . .	39
3.5	Appendix . . . . .	43
3.5.1	Consistency of contrast estimates in noise-free networks . . . . .	43
3.5.2	Standard estimators in noisy networks . . . . .	44
3.5.3	The exposure probabilities in the generalized four-level exposure model . . . . .	45
<b>4</b>	<b>Estimation of local time-varying reproduction numbers in noisy surveillance data</b>	<b>46</b>
4.1	Introduction . . . . .	47
4.2	Methods . . . . .	48
4.2.1	Notation . . . . .	49
4.2.2	Bias of the noisy local time-varying reproduction number . . . . .	50
4.2.3	Bayesian hierarchical modeling to account for misidentification . . . . .	51
4.2.4	Estimating misidentification rates . . . . .	53
4.3	Results . . . . .	54
4.3.1	Simulation study . . . . .	54
4.3.2	Application . . . . .	56
<b>5</b>	<b>Discussion</b>	<b>61</b>
5.1	Inference for metrics for network-based epidemic surveillance . . . . .	61
5.2	Experiments on noisy networks . . . . .	63

5.3	Estimation of reproduction numbers from noisy surveillance data . . .	65
<b>A</b>	<b>Supplementary Materials to Chapter 2</b>	<b>68</b>
A.1	Theorems and corollaries for bias of the observed branching number .	68
A.2	Theorems and corollaries for variance of the observed branching number	71
A.3	Proofs of theorems for bias of the observed branching number . . . .	73
A.4	Proofs of corollaries for bias of the observed branching number . . . .	76
A.5	Proofs of theorems for variances of the observed branching number .	81
A.6	Proofs of corollaries for variances of the observed branching number .	85
A.7	Proofs of theorem for the method-of-moments estimator . . . . .	99
A.8	Algorithm for estimation of asymptotic variance of method-of-moments estimator . . . . .	100
<b>B</b>	<b>Supplementary Materials to Chapter 3</b>	<b>102</b>
B.1	Proofs of propositions for noise-free networks . . . . .	102
B.2	Proofs of propositions for noisy networks . . . . .	107
B.3	Proofs of theorems for noisy networks . . . . .	111
B.4	Proofs of theorems in the generalized four-level exposure model . . .	123
B.5	Proofs for Pareto degree distribution without a cutoff . . . . .	125
	<b>References</b>	<b>127</b>
	<b>Curriculum Vitae</b>	<b>136</b>

# List of Tables

2.1	Point estimates and 95% confidence intervals for $\alpha$ and $\beta$ in the hospital and four schools. . . . .	20
3.1	The asymptotic biases of $\tilde{y}_{A\&S,i}(c_k)$ for high (top row) and low (bottom row) degree nodes. . . . .	34

# List of Figures

2.1	Biases and variances of observed branching factors in homogeneous and inhomogeneous networks with different average degrees. Error bars are 95% confidence intervals (and often not visible, due to the scale of bias versus variance). . . . .	14
2.2	Mean absolute errors (MAE) of $\hat{\kappa}$ , and 95% confidence intervals for $\kappa$ in the simulation with 500 replications for noisy networks in the hospital and schools. Reported in the plots are the relative frequencies (RF) of the event that a confidence interval covers the corresponding true value, and also the average Length of the intervals. . . . .	19
2.3	The point estimates and 95% confidence intervals for $\kappa$ in the hospital and four schools and the observed branching factor $\tilde{\kappa}$ in each round/day.	21
2.4	The point estimates and 95% confidence intervals for $R_0$ in the hospital and four schools. . . . .	22
3.1	Biases of estimators for $\bar{y}(\cdot)$ for noisy networks in four schools. Error bars are 95% confidence intervals. . . . .	41
3.2	Standard deviations of estimators for $\bar{y}(\cdot)$ for noisy networks in four schools. Error bars are 95% confidence intervals. . . . .	42
4.1	The means of daily local and imported diagnosed counts in 1,000 simulation trials for epidemics in Hong Kong and Victoria. . . . .	55

4.2	Estimations of local time-varying reproduction numbers in simulated epidemics for Hong Kong and Victoria under three sets of error misidentification rates: $\alpha_0 \sim \text{Beta}(2, 18)$ , and $\alpha_1 \sim \text{Beta}(2, 8)$ , $\text{Beta}(4, 8)$ , or $\text{Beta}(8, 8)$ . The error bands are the averages of 95% credible intervals over 1,000 trials. Note that the differences between the blue curve ( $R_*^{\text{local}}(t)$ ) and the purple curve ( $R^{\text{local}}(t)$ ) are due to the differences among infected dates, symptom onset dates, diagnosed dates. . . . .	57
4.3	Estimations of local time-varying reproduction numbers in simulated epidemics for Hong Kong and Victoria under three sets of error misidentification rates: $\alpha_1 \sim \text{Beta}(2, 18)$ , and $\alpha_0 \sim \text{Beta}(2, 8)$ , $\text{Beta}(4, 8)$ , or $\text{Beta}(8, 8)$ . The error bands are the averages of 95% credible intervals over 1,000 trials. . . . .	58
4.4	Epidemic curves of COVID-19 cases and estimations of local time-varying reproduction numbers in Hong Kong and Victoria. (a) The epidemic curve of daily cases of laboratory-confirmed SARS-CoV-2 infection in Hong Kong by symptom onset date and colored by case category. Asymptomatic cases are included here by date of confirmation. (b) The epidemic curve of the coronavirus disease cases in Victoria by sample collection date and colored by case category. (c) and (d) Estimations of local time-varying reproduction numbers under three assumed scenarios: 1) no identification error, 2) $\alpha_0 \sim \text{Beta}(2, 18)$ and $\alpha_1 \sim \text{Beta}(4, 8)$ (around 10% imported cases are misclassified as local and around 33.3% local cases are misclassified as imported), 3) $\alpha_0 \sim \text{Beta}(4, 8)$ and $\alpha_1 \sim \text{Beta}(2, 18)$ (around 33.3% imported cases are misclassified as local and around 10% local cases are misclassified as imported). The bands are the 95% credible intervals. . . . .	59

## List of Abbreviations

MAE	.....	Mean absolute errors
MCMC	.....	Markov Chain Monte Carlo
MME	.....	Method-of-moments estimators
RF	.....	Relative frequency
SEIR	.....	Susceptible-exposed-infectious-removed
SUTVA	.....	Stable unit treatment value assumption

## Chapter 1

# Introduction

The analysis of network data is widespread across the scientific disciplines. Technological and infrastructure, social, biological, and information networks are a few of the major network classes in which such analyses have been employed. Certainly, in most applied settings it is widely recognized by practitioners that there is measurement error associated with common types of network constructions. But there has been almost no serious attention to date given to the formal probabilistic characterization of the propagation of such error and statistical methods accounting for such propagation. In dissertation, I will show my work on three topics related to uncertainty analysis for noisy networks.

Chapter 2 discusses the epidemic branching factor in noisy contact networks. Epidemic modeling, while not at all new, has taken on renewed importance this year due to the COVID-19. Many key concepts in mathematical epidemiology depend on the branching factor – for example, the basic reproduction number  $R_0$ . The branching factor captures a notion of the average degree of the vertex reached by following an edge from a vertex and, therefore, measures the rate of spreading across the network. Moreover, contact network information generally is available only up to some level of error. We quantify how such errors propagate to the estimation of the branching factor, and provide a method-of-moments estimator for the true branching factor when as few as three replicates of the observed network are available. Numerical simulation suggests that high accuracy is possible for estimating branching factors in networks of

even modest size.

Chapter 3 focuses on causal inference under network interference with noise. Increasingly, there is a marked interest in estimating causal effects under network interference due to the fact that interference manifests naturally in networked experiments. Extensive work regarding uncertainty analysis has been done in causal inference without the network structure or interference. But, to our best knowledge, there has been little attention to date given towards uncertainty analysis of estimators for average causal effects under network interference. We study the propagation of such errors to estimators of average causal effects under network interference. Specifically, assuming a four-level exposure model and Bernoulli random assignment of treatment, we characterize the impact of network noise on the bias and variance of standard estimators. In addition, we propose method-of-moments estimators for bias reduction. We illustrate the practical performance of our estimators through simulation studies in British secondary school contact networks.

Chapter 4 provides estimation of local time-varying reproduction numbers in noisy surveillance data. A valuable metric in understanding infectious disease local dynamics is the local time-varying reproduction number, i.e. the expected number of secondary local cases caused by each infected individual. Accurate estimation of this quantity requires distinguishing cases arising from local transmission from those imported from elsewhere. Realistically, we can expect identification of cases as local or imported to be imperfect. We study the propagation of such errors in estimation of the local time-varying reproduction number. In addition, we propose a Bayesian framework for estimation of the true local time-varying reproduction number when identification errors exist. And we illustrate the practical performance of our estimator through simulation studies and with outbreaks of COVID-19 in Hong Kong and Victoria, Australia.

## Chapter 2

# Estimation of the Epidemic Branching Factor in Noisy Contact Networks

In this chapter, we study the propagation of network noise to the estimation of the branching factor. Specifically, we characterize the impact of network noise on the bias and variance of the observed branching factor for arbitrary true networks, with examples in sparse, dense, homogeneous and inhomogeneous networks. In addition, we propose a method-of-moments estimator for the true branching factor. We illustrate the practical performance of our estimator through simulation studies and with contact networks observed in British secondary schools and a French hospital. This chapter is adapted from Li et al. (2020a).

The organization of this chapter is as follows. Section 2.1 introduces the problem and previous relevant work. In Section 2.2 we provide background on the noise model and branching factor. In Section 2.3 we then present results for the bias and variance of the observed branching factor in sparse, dense, homogeneous and inhomogeneous networks. Section 2.4 proposes our method-of-moments estimator for the true branching factor. Numerical illustration is reported in Section 2.5. All proofs are relegated to supplementary material A.

## 2.1 Introduction

Epidemic modeling, while not at all new, has taken on renewed importance this year due to the COVID-19. Many key concepts in mathematical epidemiology depend

on the branching factor – for example, the basic reproduction number  $R_0$ . The latter is generally defined as the number of secondary infections expected in the early stages of an epidemic by a single infective in a population of susceptibles (Anderson and May (1991); Diekmann and Heesterbeek (2000)). The importance of  $R_0$  in the study of epidemics arises from its role in so-called threshold theorems, which state under which conditions the presence of an infective individual in a population will lead to an epidemic (Whittle (1955)). In the so-called configuration susceptible-exposed-infectious-removed (SEIR) models,  $R_0$  can be shown to equal  $\theta(\kappa - 1)/(\theta + \gamma)$ . Here  $\theta$  and  $\gamma$  are infection and recovery rates, respectively (Trapman et al. (2016)), while the branching factor,  $\kappa$ , is a measure of heterogeneity of a network. The branching factor captures a notion of the average degree of the vertex reached by following an edge from a vertex and, therefore, measures the rate of spreading across the network. It is evident that knowing the value of  $\kappa$  is vital for effective control responses in the early stages of an epidemic. In addition, various thresholds in epidemiological and percolation theory rely on the branching factor. In the discussion section, we provide details on how knowledge of the branching factor informs those statistics.

Increasingly, contact networks are playing an important role in the study of epidemiology. Knowledge of the structure of the network allows models to take into account individual-level behavioral heterogeneities and shifts. Network-based approaches have been explored for investigating disease outbreaks in human (Eubank et al. (2004)), livestock (Kao et al. (2006)) and wildlife (Craft et al. (2009)) populations. Moreover, contact network information generally is available only up to some level of error – also known as network noise. For example, there is often measurement error associated with network constructions, where, by ‘measurement error’ we will mean true edges being observed as non-edges, and vice versa. Such edge noise occurs in self-reported contact networks where participants may not perceive and recall all

contacts correctly (Smieszek et al. (2012)). It can also be found in sensor-based contact networks where automated proximity loggers are used to report frequency and duration of contacts (Drewe et al. (2012)). Contact tracing, and the contact networks that result, currently is playing a central role in the fight to control COVID-19 globally (especially in conjunction with testing) (Cevik et al. (2020), Juneau et al. (2020), Kretzschmar et al. (2020)). We investigate how network noise impacts on the observed value of  $\kappa$  and, therefore, on our understanding of infectious diseases spreading.

Extensive work regarding uncertainty quantification has been done in the field of non-network epidemic modeling, where populations are assumed uniform and with homogeneous mixing. Given adequate data, estimates of model parameters, such as  $\theta$  and  $\gamma$ , can be produced with accompanying standard errors. Methods for this purpose are reviewed in (Andersson and Britton, 2012, Chapter 9–12) and Becker and Britton (1999). Many studies have explored the effects of uncertainty in parameter estimation on basic epidemic quantities. For instance, there have been efforts to quantify uncertainty in  $R_0$  around recent high profile emergent events, including severe acute respiratory syndrome (SARS) (Chowell et al. (2004a)), the new influenza A (H1N1) (White et al. (2009)), and Ebola (Chowell et al. (2004b)). But, to our best knowledge, there has been little attention to date given towards uncertainty analysis of  $\kappa$  and relevant quantities in network-based epidemic models. Exceptions include real-time estimation of  $R_0$  at an early stage of an outbreak by considering the heterogeneity in contact networks (Davoudi et al. (2012)), and measurability of  $R_0$  in highly detailed sociodemographic data with the clustered contact structure assumed of the population (Liu et al. (2018)).

As remarked above, there appears to be little in the way of a formal and general treatment of the error propagation problem in network-based epidemic models. However, there are several areas in which the probabilistic or statistical treatment of

uncertainty enters prominently in network analysis. Model-based approaches include statistical methodology for predicting network topology or attributes with models that explicitly include a component for network noise (Jiang et al. (2011), Jiang and Kolaczyk (2012)), the ‘denoising’ of noisy networks (Chatterjee et al. (2015)), the adaptation of methods for vertex classification using networks observed with errors (Priebe et al. (2015)), and a general Bayesian framework for reconstructing networks from observational data (Young et al. (2020)). The other common approach to network noise is based on a ‘signal plus noise’ perspective. For example, Balachandran et al. (2017) introduced a simple model for noisy networks that, conditional on some true underlying network, assumed we observed a version of that network corrupted by an independent random noise that effectively flips the status of (non)edges. Later, Chang et al. (2020) developed method-of-moments estimators for the underlying rates of error when replicates of the observed network are available. In a somewhat different direction, uncertainty in network construction due to sampling has also been studied in some depth. See, for example, (Kolaczyk, 2009, Chapter 5) or Ahmed et al. (2014) for surveys of this area. However, in this setting, the uncertainty arises only from sampling—the subset of vertices and edges obtained through sampling are typically assumed to be observed without error.

Our contribution is to quantify how such errors propagate to the estimation of the branching factor, and to provide estimators for  $\kappa$  when as few as three replicates of the observed network are available. Adopting the noise model proposed by Balachandran et al. (2017), we characterize the impact of network noise on the bias and variance of the observed branching factor for arbitrary true networks, and we illustrate the asymptotic behaviors of these quantities on networks for varying densities and degree distributions. Our work shows that, in general, the bias in empirical branching factors can be expected to be nontrivial and is likely to dominate the variance. Accordingly,

we propose a parametric estimator of the branching factor, motivated by Chang et al. (2020), who recently developed method-of-moments estimators for network subgraph densities and the underlying rates of error when replicates of the observed network are available. Numerical simulation suggests that high accuracy is possible for estimating branching factors in networks of even modest size. We illustrate the practical use of our estimators in the context of contact networks in British secondary schools and a French hospital, where a small number of replicates are available.

## 2.2 Background

In this section, we provide essential notation and background.

### 2.2.1 Noise model

We assume the observed graph is a noisy version of a true graph. Let  $G = (V, E)$  be an undirected graph and  $G^{\text{obs}} = (V, E^{\text{obs}})$  be the observed graph, where we implicitly assume that the vertex set  $V$  is known. Denote the adjacency matrix of  $G$  by  $\mathbf{A} = (A_{i,j})_{N_v \times N_v}$  and that of  $G^{\text{obs}}$  by  $\tilde{\mathbf{A}} = (\tilde{A}_{i,j})_{N_v \times N_v}$ . Hence  $A_{i,j} = 1$  if there is a true edge between the  $i$ -th vertex and the  $j$ -th vertex, and 0 otherwise, while  $\tilde{A}_{i,j} = 1$  if an edge is observed between the  $i$ -th vertex and the  $j$ -th vertex, and 0 otherwise. And denote the degree of the  $i$ -th vertex in  $G$  and  $G^{\text{obs}}$  by  $d_i$  and  $\tilde{d}_i$ , respectively. We assume throughout that  $G$  and  $G^{\text{obs}}$  are simple.

We express the marginal distributions of the  $\tilde{A}_{i,j}$  in the form (Balachandran et al. (2017)):

$$\tilde{A}_{i,j} \sim \begin{cases} \text{Bernoulli}(\alpha_{i,j}), & \text{if } \{i, j\} \in E^c \\ \text{Bernoulli}(1 - \beta_{i,j}), & \text{if } \{i, j\} \in E, \end{cases}$$

where  $E^c = \{\{i, j\} : i, j \in V; i < j\} \setminus E$ . Drawing by analogy on the example of network construction based on hypothesis testing,  $\alpha_{i,j}$  can be interpreted as the

probability of a Type-I error on the (non)edge status for vertex pair  $\{i, j\} \in E^c$ , while  $\beta_{i,j}$  is interpreted as the probability of Type-II error, for vertex pair  $\{i, j\} \in E$ .

Our interest is in characterizing the manner in which the uncertainty in the  $\tilde{A}_{i,j}$  (as a noisy version of  $A_{ij}$ ) propagates to the branching factor. Here we focus on a general formulation of the problem in which we make the following three assumptions.

**Assumption 2.1 (Constant marginal error probabilities)** *Assume that  $\alpha_{i,j} = \alpha$  and  $\beta_{i,j} = \beta$  for all  $i < j$ , so the marginal error probabilities are  $\mathbb{P}(\tilde{A}_{i,j} = 0 | A_{i,j} = 1) = \beta$  and  $\mathbb{P}(\tilde{A}_{i,j} = 1 | A_{i,j} = 0) = \alpha$ .*

**Assumption 2.2 (Independent noise)** *The random variables  $\tilde{A}_{i,j}$ , for all  $i < j$ , are conditionally independent given  $A_{i,j}$ .*

**Assumption 2.3 (Large Graphs)**  $N_v \rightarrow \infty$ .

In Assumption 2.1, we assume that both  $\alpha$  and  $\beta$  remain constant over different edges. Under Assumption 2.2, the distributions of  $\tilde{d}_i$  is

$$\tilde{d}_i = \sum_{j=1}^{N_v} \tilde{A}_{j,i} \sim \text{Binomial}(N_v - 1 - d_i, \alpha) + \text{Binomial}(d_i, 1 - \beta).$$

Assumption 2.2 is not strictly necessary. See Remark 2.4 in Section 2.4. Assumption 2.3 reflects both the fact that the study of large graphs is a hallmark of modern applied work in complex networks and, accordingly, our desire to understand the asymptotic behavior of the branching factor and provide concise descriptions in terms of the bias and variance for large graphs.

**Remark 2.1** *Note that  $\alpha$  and  $\beta$  can be constants or  $o(1)$  as  $N_v \rightarrow \infty$ . For example, under Assumption 2.4, if  $\beta$  is constant and  $|E| = o(|E^c|)$ , then  $\alpha = o(1)$ . Thus,  $\alpha$  and  $\beta$  are actually  $\alpha(N_v)$  and  $\beta(N_v)$ . For notational simplicity, we omit  $N_v$ .*

In addition to the core Assumptions 2.1 – 2.3, we add a fourth assumption, upon which we will call periodically throughout the chapter when desiring to illustrate our results in the special case.

**Assumption 2.4 (Edge Unbiasedness)**  $\alpha|E^c| = \beta|E|$ , so that the expected number of observed edges equals the actual number of edges.

Our use of Assumption 2.4 reflects the understanding that a ‘good’ observation  $G^{\text{obs}}$  of the graph  $G$  should at the very least have roughly the right number of edges.

**Remark 2.2** *Assumption 2.4 cannot guarantee the unbiasedness of higher-order sub-graph counts. (Balachandran et al. (2017))*

### 2.2.2 The branching factor in network-based epidemic models

In general, the epidemic threshold of a network is the inverse of the largest eigenvalue of the adjacency matrix. Under some configuration models, the branching factor is often a good approximation of the largest eigenvalue (Pastor-Satorras et al. (2015)).

Let  $G$  be a network graph describing the contact structure among  $N_v$  elements in a population. If  $G$  derives from a so-called configuration model, as is commonly assumed in the network-based epidemic modeling literature, then the branching factor takes the following form (Buono et al. (2014)).

**Definition 2.1** *For graph  $G$  with  $N_v$  nodes, the branching factor is*

$$\kappa = \begin{cases} \frac{\sum_{i=1}^{N_v} d_i^2 / N_v}{\sum_{i=1}^{N_v} d_i / N_v} & \text{if } \sum_{i=1}^{N_v} d_i > 0 \\ 0 & \text{if } \sum_{i=1}^{N_v} d_i = 0, \end{cases}$$

where  $d_i$  is the degree of node  $i$ .

Accordingly, we denote the branching factor in the noisy network by  $\tilde{\kappa}$ . Besides the basic reproduction number,  $R_0$ , described in the introduction, there are other quantities depending on the observed branching factor. These include the percolation threshold  $1/(\tilde{\kappa} - 1)$ , the epidemic threshold  $1/(\tilde{\kappa} - 1)$ , and the immunization threshold  $1 - 1/(\lambda\tilde{\kappa})$ , where  $\lambda$  is the spreading rate (Pastor-Satorras et al. (2015)).

## 2.3 Bias and variance of the observed branching factor

In this section, we first quantify the asymptotic bias and variance of the observed branching factor for arbitrary true networks. We then show specific results for four typical classes of networks: sparse and homogeneous, sparse and inhomogeneous, dense and homogeneous, and dense and inhomogeneous. Next, we provide numerical illustrations. See supplementary material A.3–A.6 for all proofs related to the observed branching factor.

### 2.3.1 Arbitrary network topology

We present general results for the asymptotic bias and variance of the observed branching factor in arbitrary true networks.

**Theorem 2.1** *We define  $X = \sum_{i=1}^{N_v} \tilde{d}_i^2$ ,  $Y = \sum_{i=1}^{N_v} \tilde{d}_i$  and we assume  $\mathbb{E}Y > 0$ , and  $\mathbb{E}Y = \Omega(N_v)$  ( $N_v \rightarrow \infty$ ). Then, under Assumption 2.2, for any  $\eta > 0$ , we have*

$$\text{Bias}[\tilde{\kappa}] = \frac{\mathbb{E}X}{\mathbb{E}Y} - \kappa + \mathcal{O}\left(\frac{1}{(\mathbb{E}Y)^{1/(2+\eta)}} \frac{\mathbb{E}X}{\mathbb{E}Y}\right) \text{ as } N_v \rightarrow \infty.$$

**Remark 2.3** *Theorem 1 reflects the fact that, under certain assumptions,  $\mathbb{E}X/\mathbb{E}Y$  is a good approximation of  $\mathbb{E}(X/Y \cdot I_{\{Y>0\}})$ , i.e.,  $\mathbb{E}(\tilde{\kappa})$ .*

**Theorem 2.2** *Under assumptions in Theorem 2.1 and Assumption 2.1 and 2.4, for any  $\eta > 0$ , we have*

$$\text{Bias}[\tilde{\kappa}] = (2 - \alpha - \beta) \left[ \alpha(N_v - 1) + \beta - (\alpha + \beta)\kappa \right] + \mathcal{O}\left(\frac{1}{(\mathbb{E}Y)^{1/(2+\eta)}} \frac{\mathbb{E}X}{\mathbb{E}Y}\right)$$

as  $N_v \rightarrow \infty$ .

Theorem 2.1 shows the asymptotic bias of the observed branching factor in terms of the expectations of the first and second moments of the observed under Assumption 2.2. Theorem 2.2 relies on Assumptions 2.1 – 2.4 and provides a more explicit expression for the leading term of the asymptotic bias in this special case.

Under certain assumptions, we provide upper bounds for asymptotic variances of the observed branching factors and derive good approximations of asymptotic variances for arbitrary true networks. Considering that variances are important and complicated, we briefly show a main outcome here and give details in the supplementary material A.2.

We assume  $\mathbb{E}Y > 0$ ,  $\mathbb{E}Y = \Omega(N_v)$ , and  $1 - \beta = \Omega(N_v)$ . Then, under Assumption 2.1, 2.2, and 2.4, we have

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}\left(\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right]\right) \text{ as } N_v \rightarrow \infty.$$

This provides upper bounds for asymptotic variances of the observed branching factors. By making additional assumptions on the network density and degree distribution, we can obtain the order of the  $\mathcal{O}$  term and therefore the order of the variance.

### 2.3.2 Specific network topology

By making assumptions on the network density and degree distribution, we can obtain a more nuanced understanding of the limiting behavior of the observed branching factor in terms of bias and variance when the number of nodes tends towards infinity. Specifically, we consider the combinations of sparse versus dense and homogeneous versus inhomogeneous networks. By the term sparse we will mean a graph for which the average degree  $\bar{d} = \Theta(\log N_v)$ , and by dense,  $\bar{d} = \Theta(N_v^c)$ , where  $0 < c < 1$ . By the term homogeneous we mean the degrees follow a Poisson distribution, and by inhomogeneous, the degrees follow a truncated Pareto distribution.

**Corollary 2.1 (Sparse and homogeneous, dense and homogeneous)** *In the sparse homogeneous graph and dense homogeneous graph, under the assumptions in Theorem 2.2 and  $\beta = \mathcal{O}(1)$  ( $N_v \rightarrow \infty$ ), we have*

$$\text{Bias}[\tilde{\kappa}] = o(\kappa) \text{ as } N_v \rightarrow \infty.$$

**Corollary 2.2 (Sparse and inhomogeneous, dense and inhomogeneous)**

*In the sparse inhomogeneous graph and dense inhomogeneous graph, under the assumptions in Theorem 2.2 and  $\beta = \mathcal{O}(1)$  ( $N_v \rightarrow \infty$ ), we have*

$$\text{Bias}[\tilde{\kappa}] = \begin{cases} -\beta(2 - \alpha - \beta)\kappa + o(\kappa) & \text{if } 0 < \zeta \leq 2 \\ -\beta(2 - \alpha - \beta)\frac{\kappa}{(\zeta - 1)^2} + o(\kappa) & \text{if } \zeta > 2 \end{cases}$$

*as  $N_v \rightarrow \infty$ , where  $\zeta$  is the shape parameter of the truncated Pareto distribution.*

In summary, the observed branching factor is asymptotically unbiased in the homogeneous network setting, but asymptotically biased in the inhomogeneous network setting. The bias of the observed branching factor is negative which reflects the fact that the observed graph is typically more homogeneous than the true graph in the inhomogeneous setting. The bias depends on  $\alpha$ ,  $\beta$ , and  $\zeta$ , and when the shape  $\zeta > 2$ , the bias decreases as  $\zeta$  increases. The different results in the homogeneous and inhomogeneous network setting also reflect Remark 2.2 since the branching factor is related to the second-order moment.

**Theorem 2.3 (Sparse, dense, homogeneous, and inhomogeneous)** *In the combinations of sparse versus dense and homogeneous versus inhomogeneous networks, under the assumptions in Theorem 2.2 and  $\beta = \mathcal{O}(1)$  ( $N_v \rightarrow \infty$ ), we have*

$$\text{Var}[\tilde{\kappa}] = o(\text{Bias}[\tilde{\kappa}]) \text{ as } N_v \rightarrow \infty.$$

Note that the orders of the variances are asymptotically dominated by the corresponding biases for all four cases. Therefore, in noisy contact networks, bias would appear to be the primary concern for the observed branching factor. In turn, our simulation results (below) suggest that in practice this empirical bias can be quite substantial.

### 2.3.3 Simulation study

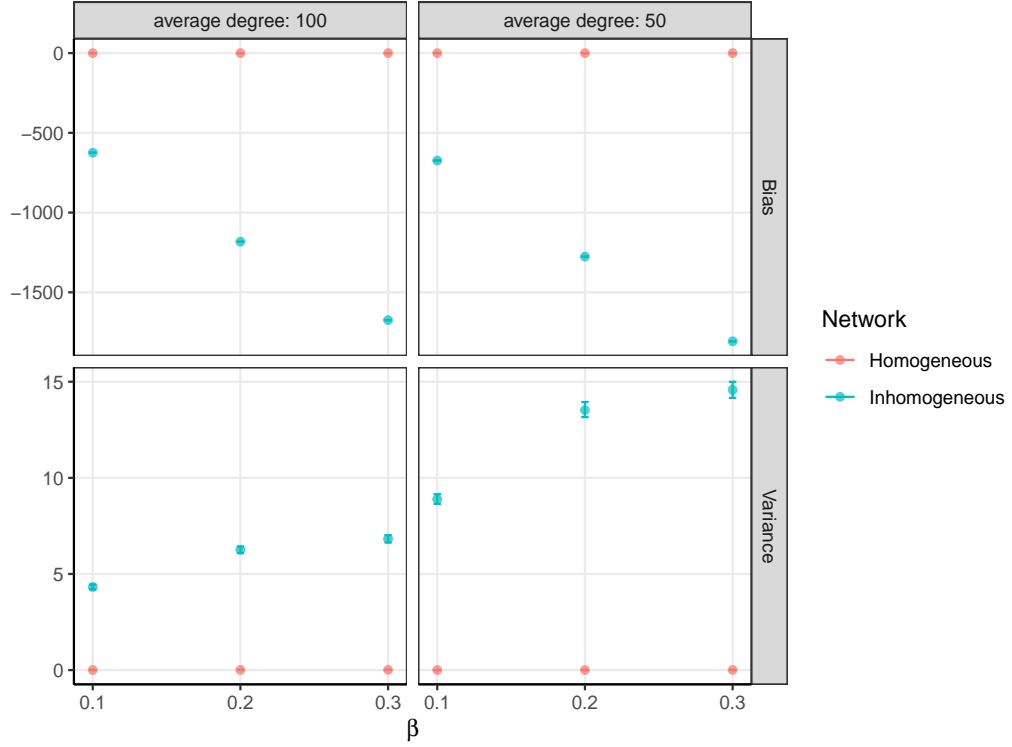
We focus on two types of networks in the simulation study: random Erdős-Rényi networks and random scale-free networks using a preferential attachment mechanism. The first type has a Poisson degree distribution, and the second type has a power law distribution. We construct two types of networks with 10,000 nodes and average degree around 50 or 100 and view them as true networks. Then we generate 10,000 noisy, observed networks according to (2.2.1). We set  $\beta = 0.1, 0.2, 0.3$  and  $\alpha = \beta|E|/|E^c|$  (i.e., to enforce edge-unbiasedness). For each observed network, we compute  $\tilde{\kappa}$ . Also, we run 1,000 times bootstrap resampling to obtain 95% confidence intervals for biases and variances. Biases and variances are shown in Figure 2.1. Error bars are 95% confidence intervals.

From the plots, we see that the noisy branching factor is unbiased in the homogeneous network setting, but biased in the inhomogeneous network setting. The bias of the observed branching factor is negative (i.e., the empirical branching factor generally underestimates the truth). And the bias increases when error rates increase. When the average degree increases from 50 to 100, the value of the true branching factor decreases from 3579.76 to 3356.34 and the bias decreases, which is consistent with Corollary 2.2. Also, variances are dominated by the corresponding biases in all cases.

## 2.4 Estimator for the true branching factor

As we saw in Section 2.3, the observed branching factor is biased in the inhomogeneous network setting. Due to the presence of heterogeneity in the level of connectivity of contact neighborhoods for most real-world contact network data, it is important to have new estimators for bias reduction. Simultaneous estimation of Type I and II errors,  $\alpha$  and  $\beta$ , as well as network quantities like  $\kappa$ , from a single noisy network is in general impossible (Chang et al., 2020, Thm 1). We present a method-of-moments

**Figure 2.1:** Biases and variances of observed branching factors in homogeneous and inhomogeneous networks with different average degrees. Error bars are 95% confidence intervals (and often not visible, due to the scale of bias versus variance).



estimator, which needs a minimum of three replicates.

We adapt the method-of-moments estimators (MME) of subgraph density in Chang et al. (2020), which require at least three replicates of the observed network. Let  $C_{\mathcal{V}_1}$  and  $C_{\mathcal{V}_2}$  denote the edge density and the two-stars density, respectively. Then

$$C_{\mathcal{V}_1} = \frac{1}{|\mathcal{V}_1|} \sum_{\mathbf{v}=(i_1, i'_1) \in \mathcal{V}_1} A_{i_1, i'_1}$$

and

$$C_{\mathcal{V}_2} = \frac{1}{|\mathcal{V}_2|} \sum_{\mathbf{v}=(i_1, i'_1, i_2, i'_2) \in \mathcal{V}_2} A_{i_1, i'_1} A_{i_2, i'_2},$$

where  $\mathcal{V}_1 = \{(i_1, i'_1) : i_1 < i'_1\}$  and  $\mathcal{V}_2 = \{(i_1, i'_1, i_2, i'_2) : i'_1 = i_2, i_1 \neq i_2 \neq i'_2\}$ .

Next we define

$$\begin{aligned}\hat{d} &= (N_v - 1)\hat{C}_{\mathcal{V}_1}, \\ \hat{d}^2 &= (N_v - 1)(N_v - 2)\hat{C}_{\mathcal{V}_2} + \hat{d},\end{aligned}$$

where  $\hat{C}_{\mathcal{V}_1}$  and  $\hat{C}_{\mathcal{V}_2}$  are method-of-moments estimators of  $C_{\mathcal{V}_1}$  and  $C_{\mathcal{V}_2}$ , which we will define later. Thus, our estimator of  $\kappa$  is given by:

$$\hat{\kappa} = \frac{\hat{d}^2}{\hat{d}} = (N_v - 2)\frac{\hat{C}_{\mathcal{V}_2}}{\hat{C}_{\mathcal{V}_1}} + 1. \quad (2.1)$$

**Theorem 2.4** *Under Assumptions 2.1 and 2.2,  $\hat{\kappa}$  has asymptotic normal distribution with mean  $\kappa$ .*

See supplementary material A.7 for proof of Theorem 2.4. Note that  $\hat{\kappa}$  is an asymptotically unbiased estimator for  $\kappa$ , where the asymptotics is in  $N_v^2$ , i.e., the square of the number of vertices in the network. To compute  $\hat{\kappa}$ , we first estimate  $C_{\mathcal{V}_1}$  and  $C_{\mathcal{V}_2}$  by methods used in Chang et al. (2020). Define relevant quantities as follows:

$$\begin{aligned}u_1 &= (1 - \delta)\alpha + \delta(1 - \beta), \\ u_2 &= (1 - \delta)\alpha(1 - \alpha) + \delta\beta(1 - \beta), \\ u_3 &= (1 - \delta)\alpha(1 - \alpha)^2 + \delta\beta^2(1 - \beta),\end{aligned}$$

where  $\delta$  is the edge density in the true network,  $u_1$  is the expected edge density in one observed network,  $u_2$  is the expected density of edge differences in two observed networks, and  $u_3$  is the average probability of having an edge between two arbitrary nodes in one observed network but no edge between same nodes in the other two

observed networks. The method-of-moments estimators for  $u_1$ ,  $u_2$  and  $u_3$  are

$$\begin{aligned}\hat{u}_1 &= \frac{2}{N_v(N_v - 1)} \sum_{i < j} \tilde{A}_{i,j}, \\ \hat{u}_2 &= \frac{1}{N_v(N_v - 1)} \sum_{i < j} |\tilde{A}_{i,j,*} - \tilde{A}_{i,j}|, \\ \hat{u}_3 &= \frac{2}{3N_v(N_v - 1)} \sum_{i < j} I(\text{Exactly one of } \tilde{A}_{i,j,**}, \tilde{A}_{i,j,*}, \tilde{A}_{i,j} \text{ equals 1}).\end{aligned}\tag{2.2}$$

where  $\tilde{\mathbf{A}}_* = (\tilde{A}_{i,j,*})_{N_v \times N_v}$ ,  $\tilde{\mathbf{A}}_{**} = (\tilde{A}_{i,j,**})_{N_v \times N_v}$  are independent and identically distributed replicates of  $\tilde{\mathbf{A}}$ .

Calculation of the estimator  $\hat{\kappa}$  in (2.1) and the estimation of its asymptotic variance can be accomplished as detailed in Algorithm 2.1 below and Algorithm A.1 in the supplementary material A.8, respectively. The variance estimation is based on a nonstandard bootstrap, as introduced in Chang et al. (2020).

---

**Algorithm 2.1** Method-of-moments estimator  $\hat{\kappa}$

---

**Input:**  $\tilde{\mathbf{A}} = (\tilde{A}_{i,j})_{N_v \times N_v}$ ,  $\tilde{\mathbf{A}}_* = (\tilde{A}_{i,j,*})_{N_v \times N_v}$ ,  $\tilde{\mathbf{A}}_{**} = (\tilde{A}_{i,j,**})_{N_v \times N_v}$ ,  $\alpha_0$ ,  $\varepsilon$

**Output:**  $\hat{\alpha}$ ,  $\hat{\beta}$ ,  $\hat{\kappa}$

Compute  $\hat{u}_1$ ,  $\hat{u}_2$ ,  $\hat{u}_3$  defined in (2.2);

Initialize  $\hat{\alpha} = \alpha_0$ ,  $\alpha_0 = \hat{\alpha} + 10\varepsilon$ ;

**while**  $|\hat{\alpha} - \alpha_0| > \varepsilon$  **do**

$$\alpha_0 \leftarrow \hat{\alpha}, \quad \hat{\beta} \leftarrow \frac{\hat{u}_2 - \alpha_0 + \hat{u}_1 \alpha_0}{\hat{u}_1 - \alpha_0}, \quad \hat{\delta} \leftarrow \frac{(\hat{u}_1 - \alpha_0)^2}{\hat{u}_1 - \hat{u}_2 - 2\hat{u}_1 \alpha_0 + \alpha_0^2}, \quad \hat{\alpha} \leftarrow \frac{\hat{u}_3 - \hat{\delta} \hat{\beta}^2 (1 - \hat{\beta})}{(1 - \hat{\delta})(1 - \alpha_0)^2};$$

Compute  $\hat{k}_3 = 1 - \hat{\alpha} - \hat{\beta}$ ,  $\hat{C}_{\mathcal{V}_1} = \frac{2}{\hat{k}_3 N_v (N_v - 1)} \sum_{i < j} (\tilde{A}_{i,j} - \hat{\alpha})$ ,

$$\hat{C}_{\mathcal{V}_2} = \frac{1}{\hat{k}_3^2 N_v (N_v - 1) (N_v - 2)} \sum_{i \neq j \neq l} (\tilde{A}_{i,j} - \hat{\alpha})(\tilde{A}_{j,l} - \hat{\alpha}), \quad \hat{\kappa} = (N_v - 2) \frac{\hat{C}_{\mathcal{V}_2}}{\hat{C}_{\mathcal{V}_1}} + 1.$$


---

**Remark 2.4** *Since our estimation of the unknown parameters is based on moment estimation, the independent noise dictated by Assumption 2.2 is not strictly necessary. As is shown in the proof of Chang et al. (2020), the convergence rate for the moment estimation of the unknown parameters is determined by the convergence rates of  $\hat{u}_1 - u_1$ ,  $\hat{u}_2 - u_2$  and  $\hat{u}_3 - u_3$ . When some limited dependency among observed edges is present, the convergence rates of  $\hat{u}_1 - u_1$ ,  $\hat{u}_2 - u_2$  and  $\hat{u}_3 - u_3$  still are  $\mathcal{O}(1/N_v)$ .*

## 2.5 Numerical illustration

In this section, we conduct some simulations and experiments to illustrate the finite sample properties of the proposed estimation methods. We consider two types of contact networks. One is the self-reported British secondary school contact network, described in Kucharski et al. (2018). These data were collected from 460 unique participants across four rounds of data collection conducted between January and June 2015 in year 7 groups in four UK secondary schools, with 7,315 identifiable contacts reported in total. They used a process of peer nomination as a method for data collection: students were asked, via the research questionnaire, to list the six other students in year 7 at their school that they spend the most time with. For each pair of participants in a specific round of data collection, a single link was defined if either one of the participants reported a contact between the pair (i.e. there was at least one unidirectional link, in either direction). Our analysis focuses on the single link contact network.

The other contact network we used is a sensor-based contact network in a French Hospital, reported by Vanhems et al. (2013). These data contain records of contacts among patients and various types of health care workers in the geriatric unit of a hospital in Lyon, France, in 2010, from 1pm on Monday, December 6 to 2pm on Friday, December 10. Each of the 75 people in this study consented to wear RFID sensors on small identification badges during this period, which made it possible to record when any two of them were in face-to-face contact with each other (i.e., within 1-1.5 m of each other) during a 20-second interval of time. A primary goal of this study was to gain insight into the pattern of contacts in such a hospital environment, particularly with an eye towards the manner in which infection might be transmitted. We define a link if duration of contacts in one day is greater than 5 minutes and construct networks for Tuesday, Wednesday and Thursday.

Each data set has at least three replicates. And we consider two settings, a simulation setting where noise is added to a ‘true’ network derived from the data and an application setting where three replicates are each treated as noisy versions of an unknown true network. The former results allow us to understand what finite-sample properties can be expected of our estimators, while the latter are reflective of what would be observed in practice with such data.

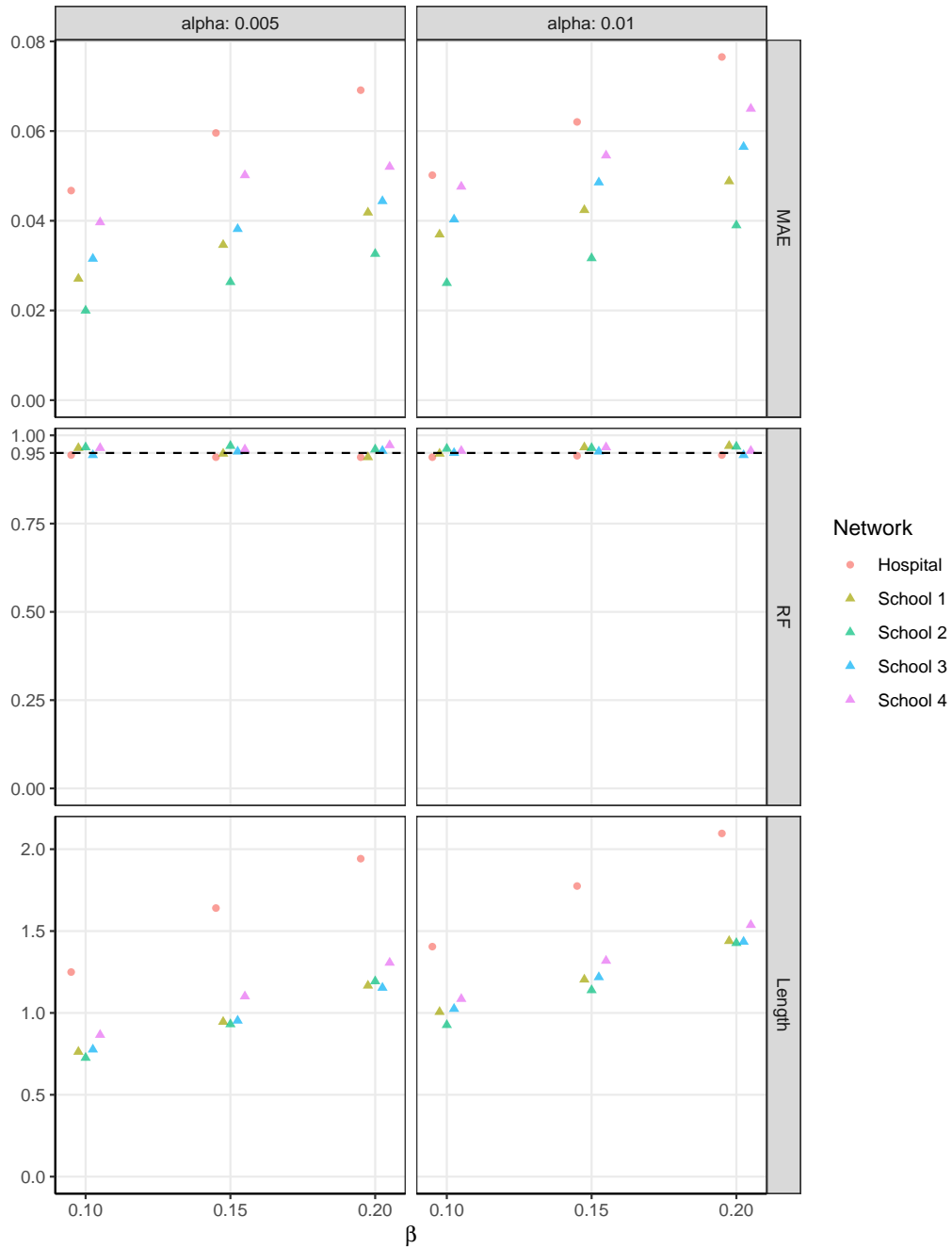
### 2.5.1 Simulations

For each data set, we artificially constructed a ‘true’ adjacency matrix  $\mathbf{A}$ : if an edge occurs between a pair of vertices more than once in observed networks, we view that pair to have a true edge. The noisy, observed adjacency matrices  $\tilde{\mathbf{A}}$ ,  $\tilde{\mathbf{A}}_*$ ,  $\tilde{\mathbf{A}}_{**}$  are generated according to (2.2.1). We set  $\alpha = 0.005$  or  $0.010$ , and  $\beta = 0.01$ ,  $0.15$ , or  $0.20$ . We assume that both  $\alpha$  and  $\beta$  are unknown.

We evaluate the method-of-moments estimate for  $\kappa$  and 95% confidence intervals. Figure 2.2 shows the simulation results, in which we replicate 500 times for each setting. The mean absolute errors (MAE) for the point estimates for the branching factor  $\kappa$  and the relative frequency (RF) of coverage for the estimated 95% confidence interval for  $\kappa$  are shown in Figure 2.2. Note that,  $\text{MAE}(\hat{\kappa}) = \frac{1}{500} \sum_{i=1}^{500} |\hat{\kappa}_i - \kappa|$ , where  $\hat{\kappa}_1, \dots, \hat{\kappa}_{500}$  denote the estimated values in 500 replications of simulation, and  $\kappa$  denotes the true value.

In the hospital and school networks, the estimation errors for  $\kappa$  increase when  $\alpha$  and  $\beta$  increase. And the estimated coverage probabilities are indeed around 95%. The average interval lengths in the French hospital are larger than that in the four schools due to smaller sample size.

**Figure 2.2:** Mean absolute errors (MAE) of  $\hat{\kappa}$ , and 95% confidence intervals for  $\kappa$  in the simulation with 500 replications for noisy networks in the hospital and schools. Reported in the plots are the relative frequencies (RF) of the event that a confidence interval covers the corresponding true value, and also the average Length of the intervals.



### 2.5.2 Application

In the school data sets, the nodes are not all the same within a given school over the four rounds. So, we choose the nodes common over four rounds and their edges to obtain four replicates of the noisy networks. Since our estimation methods only need three replicates, we select rounds 1, 2, and 3 (analogous results hold for other choices). Similarly, for the hospital data set, we choose the nodes common over three days and their edges to obtain three replicates of the noisy networks.

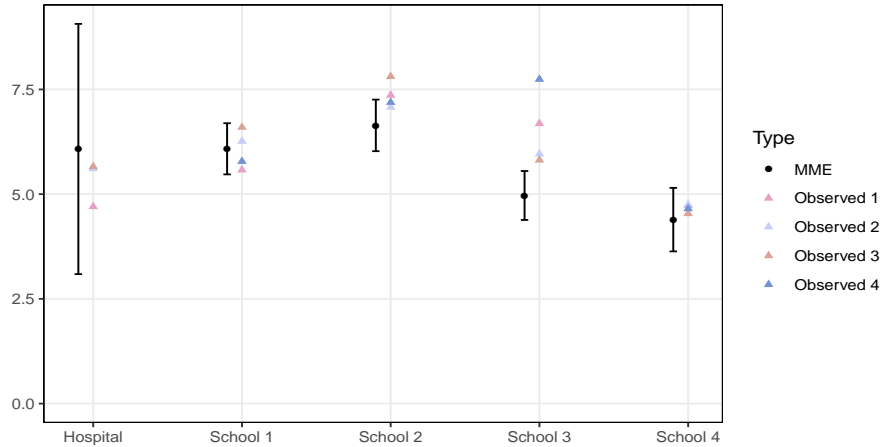
We evaluate the method-of-moments estimates for  $\kappa$ , 95% confidence intervals, and the observed branching factor  $\tilde{\kappa}$ . Point estimates and 95% confidence intervals for  $\alpha$  and  $\beta$  are reported in Table 2.1. Figure 2.3 show the point estimates for the branching factor  $\kappa$  and the observed branching factor  $\tilde{\kappa}$  in each round. The error bars are the estimated 95% confidence interval for  $\kappa$ .

**Table 2.1:** Point estimates and 95% confidence intervals for  $\alpha$  and  $\beta$  in the hospital and four schools.

Networks	Estimates	$\alpha$ CI	Estimates	$\beta$ CI
Hospital	0.116	( 0.080, 0.153 )	0.162	( -0.173, 0.499 )
School 1	0.005	( 0.004, 0.007 )	0.207	( 0.140, 0.275 )
School 2	0.013	( 0.012, 0.015 )	0.141	( 0.092, 0.191 )
School 3	0.013	( 0.012, 0.015 )	0.000	( -0.057, 0.057 )
School 4	0.020	( 0.014, 0.025 )	0.123	( 0.025, 0.222 )

Table 2.1 indicates there exists nontrivial noise in all networks. The estimate of  $\alpha$  in the hospital network is one order of magnitude larger than that in the school networks. Figure 2.3 shows that, in schools 2 and 3, the resulting method-of-moments estimates for  $\kappa$  are lower than all of their observed values, indicating a nontrivial downward adjustment for network noise. And most of the observed branching factors are not in the estimated 95% confidence intervals, which further reinforces the evidence that the true branching factor is less than those observed empirically. In schools 1 and 4, the resulting method-of-moments estimates for  $\kappa$  are close to their observed

**Figure 2.3:** The point estimates and 95% confidence intervals for  $\kappa$  in the hospital and four schools and the observed branching factor  $\tilde{\kappa}$  in each round/day.



values. In contrast, in the French hospital, the estimate for  $\kappa$  is higher than all of their observed values, indicating a nontrivial upward adjustment.

Ultimately, we see that the ability to account for network noise appropriately in reporting the branching factor can lead to substantially different conclusions than use of the original, empirically observed branching factor. These differences can then in turn be translated to specific epidemic-related quantities of interest in a study.

For example, recall that  $R_0$  equals  $\theta(\kappa - 1)/(\theta + \gamma)$  in the network-based SEIR model, where  $\theta$  and  $\gamma$  are infection and recovery rates. Therefore, if we are interested in characterizing the manner in which the uncertainty in the branching factor propagates to  $R_0$ , we can do so given knowledge or conjecture of values for these rates. Consider the context of COVID-19, for example, for which current best knowledge suggests parameter settings of  $\theta = 0.016$  or  $0.026$  and  $1/\gamma$  from 8 to 24.6 (Luo et al. (2020); Lauer et al. (2020); Linton et al. (2020); Wang et al. (2020); Wölfel et al. (2020); Verity et al. (2020)). Estimating infection and recovery rates are important in epidemic modeling, but we treat  $\theta$  and  $\gamma$  as constants here for illustration, and only consider the uncertainty in the branching factor.

**Figure 2.4:** The point estimates and 95% confidence intervals for  $R_0$  in the hospital and four schools.

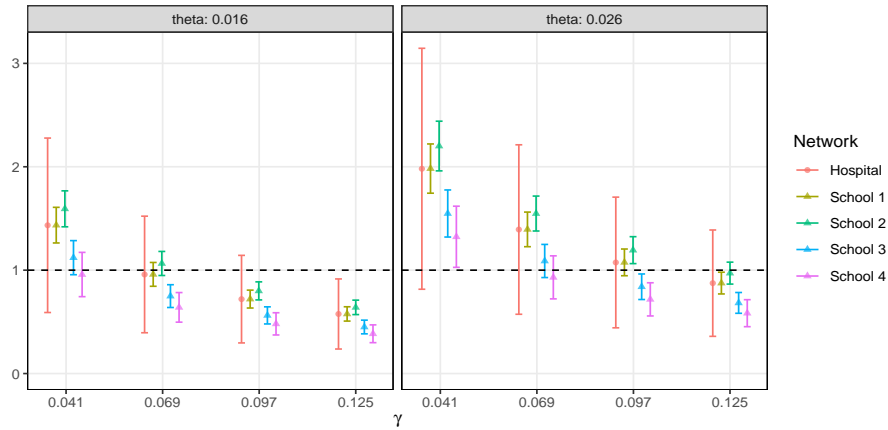


Figure 2.4 shows the point estimates and 95% confidence intervals for  $R_0$  in the hospital and four schools. School 2 consistently has the highest estimated  $\hat{R}_0$ . The infection will be able to start spreading in a population when  $R_0 > 1$ , but not if  $R_0 < 1$ . For school networks, most of the 95% confidence intervals include 1 or are below 1 when  $\theta = 0.016$ , while some are higher when  $\theta = 0.026$ . The 95% confidence intervals include 1 in all cases for the French hospital.

## Data accessibility

No primary data are used in this chapter. Secondary data sources are taken from Kucharski et al. (2018) and Vanhems et al. (2013). These data and the code necessary to reproduce the results in this chapter are available at <https://github.com/KolaczykResearch/EstimNetReprodNumber>.

## Chapter 3

# Causal Inference under Network Interference with Noise

In this chapter, we quantify biases and variances of standard estimators of average causal effects in noisy networks and develop a general framework for estimation of true average causal effects in contexts wherein one has observations of noisy networks. Our approach requires as few as three replicates of network observations, and employs method-of-moments techniques to derive estimators and establish their asymptotic unbiasedness and consistency. Simulations in British secondary schools contact networks demonstrate that substantial inferential accuracy by method-of-moments estimators is possible in networks of even modest size when nontrivial noise is present.

The organization of this chapter is as follows. Section 3.1 sets up the problem and reviews previous relevant work. In Section 3.2 we present the bias and variance of standard estimators in noisy networks under a four-level exposure model and Bernoulli random assignment of treatment. Section 3.3 contains our proposed method-of-moments estimators for the true average causal effects. Numerical illustrations are reported in Section 3.4. All proofs are relegated to supplementary material B.

### 3.1 Introduction

In recent years, there has been an enormous interest in the assessment of treatment effects within networked systems. Naturally, interference (Cox and Cox (1958)) cannot

realistically be assumed away when doing experiments on networks. The outcome of one individual may be affected by the treatment assigned to other individuals, which violates the ‘stable unit treatment value assumption’ (SUTVA) (Neyman (1923), Rubin (1990)). As a result, much of what is considered standard in the traditional design of randomized experiments and the corresponding analysis for causal inference does not apply directly in this context.

Moreover, network information generally is available only up to some level of error, also known as network noise. For example, there is often measurement error associated with network constructions, where, by ‘measurement error’ we will mean true edges being observed as non-edges, and vice versa. Such edge noise occurs in self-reported contact networks where participants may not perceive and recall all contacts correctly (Smieszek et al. (2012)). It can also be found in biological networks (e.g., of gene regulatory relationships), which are often based on notions of association (e.g., correlation, partial correlation, etc.) among experimental measurements of gene activity levels that are determined by some form of statistical inference. We investigate how network noise impacts estimators of average causal effects under network interference and how to account for the noise.

### 3.1.1 Problem setup

We assume the observed graph is a noisy version of a true graph. Let  $G = (V, E)$  be an undirected graph and  $G^{\text{obs}} = (V, E^{\text{obs}})$  be the observed graph, where we assume that the vertex set  $V$  is known. Denote the adjacency matrix of  $G$  by  $\mathbf{A} = (A_{i,j})_{N_v \times N_v}$  and that of  $G^{\text{obs}}$  by  $\tilde{\mathbf{A}} = (\tilde{A}_{i,j})_{N_v \times N_v}$ . Hence  $A_{i,j} = 1$  if there is a true edge between the  $i$ -th vertex and the  $j$ -th vertex, and 0 otherwise, while  $\tilde{A}_{i,j} = 1$  if an edge is observed between the  $i$ -th vertex and the  $j$ -th vertex, and 0 otherwise. We assume throughout that  $G$  and  $G^{\text{obs}}$  are simple.

We express the marginal distributions of the  $\tilde{A}_{i,j}$  in the form (Balachandran et al.

(2017)):

$$\tilde{A}_{i,j} \sim \begin{cases} \text{Bernoulli}(\alpha_{i,j}), & \text{if } \{i,j\} \in E^c, \\ \text{Bernoulli}(1 - \beta_{i,j}), & \text{if } \{i,j\} \in E, \end{cases} \quad (3.1)$$

where  $E^c = \{\{i,j\} : i,j \in V; i < j\} \setminus E$ . Drawing by analogy on the example of network construction based on hypothesis testing,  $\alpha_{i,j}$  can be interpreted as the probability of a Type-I error on the (non)edge status for vertex pair  $\{i,j\} \in E^c$ , while  $\beta_{i,j}$  is interpreted as the probability of Type-II error, for vertex pair  $\{i,j\} \in E$ . Our interest is in characterizing the manner in which the uncertainty in the  $\tilde{A}_{i,j}$  propagates to estimators of average causal effects.

Let  $z_i = 1$  indicate that individual  $i \in V$  received a given treatment. We will refer to  $\mathbf{z} = (z_1, \dots, z_{N_v})^\top \in \{0, 1\}^{N_v}$  as the treatment assignment vector. Let  $p_{\mathbf{z}} = \mathbb{P}(\mathbf{Z} = \mathbf{z})$  be the probability that treatment assignment  $\mathbf{z}$  is generated by the experimental design. Additionally, let  $y_i(\mathbf{z})$  denote the outcome for individual  $i$  under treatment assignment  $\mathbf{z}$ . In the worst case, there will be  $2^{N_v}$  possible exposures for each of the  $N_v$  individuals, making causal inference impossible. To avoid this situation, we adopt the notion of so-called exposure mappings, introduced by Aronow and Samii (2017). We say that  $i$  is exposed to condition  $k = 1, \dots, K$  if  $f(\mathbf{z}, \mathbf{x}_i) = c_k$ , where  $f$  is the exposure mapping,  $\mathbf{z}$  is the treatment assignment vector, and  $\mathbf{x}_i$  is a vector of additional information specific to individual  $i$ . Under interference, these authors offer a simple, four-level categorization of exposure ( $K = 4$ ) that we revisit here and throughout this chapter. Taking the vector  $\mathbf{x}_i$  to be the  $i$ th column of the adjacency

matrix  $\mathbf{A}$  (i.e.,  $\mathbf{x}_i = \mathbf{A}_i$ ), they define

$$f(\mathbf{z}, \mathbf{A}_i) = \begin{cases} c_{11}(\text{Direct} + \text{Indirect Exposure}), & z_i I_{\{\mathbf{z}^\top \mathbf{A}_i > 0\}} = 1, \\ c_{10}(\text{Isolated Direct Exposure}), & z_i I_{\{\mathbf{z}^\top \mathbf{A}_i = 0\}} = 1, \\ c_{01}(\text{Indirect Exposure}), & (1 - z_i) I_{\{\mathbf{z}^\top \mathbf{A}_i > 0\}} = 1, \\ c_{00}(\text{No Exposure}), & (1 - z_i) I_{\{\mathbf{z}^\top \mathbf{A}_i = 0\}} = 1, \end{cases} \quad (3.2)$$

where the inner product  $\mathbf{z}^\top \mathbf{A}_i$  is the number of treated neighbors of individual  $i$ .

In the general exposure mapping framework of Aronow and Samii (2017), potential outcomes are dependent only on the exposure conditions for each unit. Suppose each individual  $i$  has  $K$  potential outcomes  $y_i(c_1), \dots, y_i(c_K)$  and is exposed to one and only one condition. Then, define

$$\tau(c_k, c_l) = \frac{1}{N_v} \sum_{i=1}^{N_v} [y_i(c_k) - y_i(c_l)] = \bar{y}(c_k) - \bar{y}(c_l) \quad (3.3)$$

to be the average causal contrast between exposure condition  $k$  versus  $l$ . Consider again, for example, the exposure mapping function defined in (3.2). A natural set of contrasts is  $\tau(c_{01}, c_{00})$ ,  $\tau(c_{10}, c_{00})$ , and  $\tau(c_{11}, c_{00})$ , which capture the average indirect treatment effect, the average direct treatment effect, and the average total treatment effect, respectively.

Now consider the problem of inference for causal effects under network interference. The Horvitz-Thompson framework accounts for unequal-probability sampling through the use of inverse probability weighting (Horvitz and Thompson (1952)) and is adapted by Aronow and Samii (2017) under exposure mappings. In noise-free networks, assuming all individuals have nonzero exposure probabilities for all exposure conditions, the estimator

$$\hat{y}(c_k) = \frac{1}{N_v} \left\{ \sum_{i=1}^{N_v} I_{\{f(\mathbf{Z}, \mathbf{x}_i)=c_k\}} \frac{y_i(c_k)}{p_i^e(c_k)} \right\} \quad (3.4)$$

is well-defined and unbiased for  $\bar{y}(c_k)$ , where the exposure probabilities  $p_i^e(c_k)$  are defined as  $\sum_{\mathbf{z}} p_{\mathbf{z}} I_{\{f(\mathbf{z}, \mathbf{x}_i)=c_k\}}$ . In turn,  $\hat{\tau}(c_k, c_l) = \hat{y}(c_k) - \hat{y}(c_l)$  is an unbiased estimator of  $\tau(c_k, c_l)$ .

However, in noisy networks, exposure levels will be misclassified. For example, in the four-level exposure model, for a node  $i$ , the expected confusion matrix for observed (rows) versus true (columns) exposures has the following form

$$\mathbf{P}_i := \begin{bmatrix} P_i(\tilde{c}_{11}, c_{11}) & P_i(\tilde{c}_{11}, c_{10}) & 0 & 0 \\ P_i(\tilde{c}_{10}, c_{11}) & P_i(\tilde{c}_{10}, c_{10}) & 0 & 0 \\ 0 & 0 & P_i(\tilde{c}_{01}, c_{01}) & P_i(\tilde{c}_{01}, c_{00}) \\ 0 & 0 & P_i(\tilde{c}_{00}, c_{01}) & P_i(\tilde{c}_{00}, c_{00}) \end{bmatrix}, \quad (3.5)$$

where  $\tilde{c}_k$  represents the exposure level in observed networks and  $P_i(\tilde{c}_k, c_l) = \mathbb{E}[I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i)=\tilde{c}_k\}} I_{\{f(\mathbf{Z}, \mathbf{A}_i)=c_l\}}]$ . The two off-diagonal blocks are equal to 0, since network noise does not affect treatment status. The four symbols  $P_i(\tilde{c}_k, c_l), k \neq l$  are cases where exposure levels are misclassified. In the general exposure mapping framework, the estimators (3.4) for  $\bar{y}(c_k)$  are in fact

$$\tilde{y}_{A\&S}(c_k) = \frac{1}{N_v} \sum_{i=1}^{N_v} I_{\{\tilde{p}_i^e(c_k) > 0\}} I_{\{f(\mathbf{Z}, \tilde{\mathbf{X}}_i)=c_k\}} \frac{1}{\tilde{p}_i^e(c_k)} \left\{ \sum_{l=1}^K y_i(c_l) I_{\{f(\mathbf{Z}, \mathbf{x}_i)=c_l\}} \right\}, \quad (3.6)$$

where  $\tilde{\mathbf{X}}_i$  is a noisy version of  $\mathbf{x}_i$ , and  $\tilde{p}_i^e(c_k) = \sum_{\mathbf{z}} p_{\mathbf{z}} I_{\{f(\mathbf{z}, \tilde{\mathbf{X}}_i)=c_k\}}$ . From (3.6), we can see that the errors come from two parts: incorrect exposure probabilities and misclassified exposure levels.

In this chapter, we will address the following important questions. First, what is the impact of ignoring network noise? Second, how can we account for network noise?

### 3.1.2 Related literature

Awareness of interference goes back at least 100 years (e.g., Ross (1916)), and its impact on standard theory and methods has been studied previously in certain specific contexts, including interference localized to an individual across different rounds of treatment in clinical trials with crossover designs (Grizzle (1965)), interference based on spatial proximity of treated units (Kempton and Lockwood (1984)) and interference within blocks (Hudgens and Halloran (2008)). For network interference, an assumption that has gained traction is that the causal effects can be passed along edges in the network. A highly studied assumption is to assume that unit outcomes are only impacted by their neighbors in the network (Manski (2013); Athey et al. (2018)). Researchers have recently developed frameworks for estimating average unit-level causal effects under network interference. For example, Aronow and Samii (2017) provided unbiased estimators of average unit-level causal effects induced by treatment exposure. Sussman and Airoidi (2017) proposed minimum integrated variance linear unbiased estimators with respect to a distribution on the potential outcomes.

Extensive work regarding uncertainty analysis has been done in causal inference without the network structure or interference. Many studies have explored the effects of uncertainty in propensity scores on causal inference. For instance, there have been efforts to develop Bayesian propensity score estimators to incorporate such uncertainties into causal inference (e.g., An (2010), Alvarez and Levin (2014)). And there are some studies on the properties for particular matching estimators for average causal effects (e.g., Abadie and Imbens (2006), Schafer and Kang (2008)). But, to our best knowledge, there has been little attention to date given towards uncertainty analysis of estimators for average causal effects under network interference. Exceptions include a Bayesian procedure which accounts for network uncertainty and relies on a linear response assumption to increase estimation precision (Toulis and Kao (2013)),

and structure learning techniques to estimate causal effects under data dependence induced by a network represented by a chain graph model, when the structure of this dependence is not known a priori (Bhattacharya et al. (2019)).

As remarked above, there appears to be little in the way of a formal and general treatment of the error propagation problem in estimators of average causal effects under network interference. However, there are several areas in which the probabilistic or statistical treatment of uncertainty enters prominently in network analysis. Model-based approaches include statistical methodology for predicting network topology or attributes with models that explicitly include a component for network noise (Jiang et al. (2011), Jiang and Kolaczyk (2012)), the ‘denoising’ of noisy networks (Chatterjee et al. (2015)), the adaptation of methods for vertex classification using networks observed with errors (Priebe et al. (2015)), a regression model on network-linked data that is based on a flexible network effect assumption and is robust to errors in the network structure (Le and Li (2020)), and a general Bayesian framework for reconstructing networks from observational data (Young et al. (2020)). The other common approach to network noise is based on a ‘signal plus noise’ perspective. For example, Balachandran et al. (2017) introduced a simple model for noisy networks that, conditional on some true underlying network, assumed we observed a version of that network corrupted by an independent random noise that effectively flips the status of (non)edges. Later, Chang et al. (2020) developed method-of-moments estimators for the underlying rates of error when replicates of the observed network are available. In a somewhat different direction, uncertainty in network construction due to sampling has also been studied in some depth. See, for example, (Kolaczyk, 2009, Chapter 5) or Ahmed et al. (2014) for surveys of this area. However, in that setting, the uncertainty arises only from sampling—the subset of vertices and edges obtained through sampling are typically assumed to be observed without error.

### 3.1.3 Our contributions

Our contribution in this chapter is to quantify how network errors propagate to standard estimators of average causal effects under network interference, and to provide new estimators for average causal effects when replicates of the observed network are available. Adopting the noise model proposed by Balachandran et al. (2017), we characterize the impact of network noise on the bias and variance of standard estimators (Aronow and Samii (2017)) under a four-level exposure model and Bernoulli random assignment of treatment, and we illustrate the asymptotic behaviors on networks for varying degree distributions. Additionally, we propose method-of-moments estimators of average causal effects, when replicates of the observed network are available. Numerical simulation in the context of social contact networks in British secondary schools suggests that high accuracy is possible for networks of even modest size.

## 3.2 Impact of ignoring network noise

In this section, we characterize the impact of network noise on biases and variances of standard estimators under a four-level exposure model and Bernoulli random assignment of treatment. Specifically, we show results for two typical classes of networks: homogeneous and inhomogeneous. By the term homogeneous we mean the degrees follow a zero-truncated Poisson distribution, and by inhomogeneous, the degrees follow a Pareto distribution with an exponential cutoff (Clauset et al. (2009)). Note that many real networks present a bounded scale-free behavior with a connectivity cut-off due to the finite size of the network or to the presence of constraints limiting the addition of new links in an otherwise infinite network (Amaral et al. (2000)). The exponential cutoff is most widely used.

### 3.2.1 Network settings and assumptions

We consider two typical classes of networks: homogeneous and inhomogeneous. The formal definitions are as follows.

**Homogeneous network setting** The degree distribution of  $G$  is a zero-truncated Poisson distribution with mean  $\bar{d}$ .

**Inhomogeneous network setting** The degree distribution of  $G$  is a Pareto distribution with an exponential cutoff with rate  $\lambda$ , shape  $\zeta$ , lower bound  $d_L$ , upper bound  $N_v - 1$  and mean  $\bar{d}$ .

**Remark 3.1** *The degree distribution is the probability distribution of the degrees over the whole network.*

**Remark 3.2** *Note that  $\bar{d}$ ,  $\lambda$  and  $d_L$  depend on  $N_v$ . For notational simplicity, we omit  $N_v$ .*

**Remark 3.3** *In the inhomogeneous network setting, by the definition of Pareto distribution with an exponential cutoff,  $\lambda$ ,  $\zeta$ ,  $d_L$ ,  $\bar{d}$  and  $N_v$  satisfy the equation*

$$\bar{d} = \int_{d_L}^{N_v-1} x \cdot e^{-\lambda x} x^{-(\zeta+1)} dx \Big/ \int_{d_L}^{N_v-1} e^{-\lambda x} x^{-(\zeta+1)} dx.$$

Here we focus on a general formulation of the problem in which we make the following assumptions on networks and the treatment assignment.

**Assumption 3.1 (Constant marginal error probabilities)** *Assume that  $\alpha_{i,j} = \alpha$  and  $\beta_{i,j} = \beta$  for all  $i < j$ , so the marginal error probabilities are  $\mathbb{P}(\tilde{A}_{i,j} = 0 | A_{i,j} = 1) = \beta$  and  $\mathbb{P}(\tilde{A}_{i,j} = 1 | A_{i,j} = 0) = \alpha$ .*

**Assumption 3.2 (Independent noise)** *The random variables  $\tilde{A}_{i,j}$ , for all  $i < j$ , are conditionally independent given  $A_{i,j}$ .*

**Assumption 3.3 (Large Graphs)** *The number of vertices  $N_v \rightarrow \infty$ .*

In Assumption 3.1, we assume that both  $\alpha$  and  $\beta$  remain constant over different edges. Under Assumptions 3.1 and 3.2, the distribution of  $\tilde{d}_i$  is

$$\tilde{d}_i = \sum_{j=1}^{N_v} \tilde{A}_{j,i} \sim \text{Binomial}(N_v - 1 - d_i, \alpha) + \text{Binomial}(d_i, 1 - \beta).$$

Assumption 3.3 reflects both the fact that the study of large graphs is a hallmark of modern applied work in complex networks and, accordingly, our desire to understand asymptotic behaviors of estimators for average causal effects and provide concise descriptions in terms of biases and variances for large graphs.

**Assumption 3.4** *The treatment probability  $p$  satisfies  $p = o(1)$ ,  $p = \omega(1/N_v)$ ,  $\bar{d} = \Theta(1/p)$ . Letting  $C_{ij}$  denote the number of common neighbors between vertices  $i$  and  $j$  in  $G$ ,  $\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} I_{\{C_{ij}=0\}} \sim N_v^2$ . Finally, the potential outcomes are bounded,  $|y_i(c_k)| \leq c < \infty$ , for all values  $i$  and  $c_k$ , where  $c$  is a constant.*

Assumption 3.4 entails that, as  $N_v$  grows, the expected number of treated individuals also grows but is dominated by  $N_v$  asymptotically. And the average number of treated neighbors is bounded. The amount of vertex pairs having common neighbors is also limited in scope as  $N_v$  grows which ensures a sufficiently large set of independent exposures. Assumption 3.4 is an assumption used in proving the consistency of  $\hat{\tau}(c_k, c_l)$  in noise-free homogeneous and inhomogeneous networks. See Appendix 3.5.1 for details.

**Assumption 3.5**  $1 - \beta = \Omega(1)$ ,  $\alpha = \Theta(1/(pN_v))$ , and  $\alpha = o(p)$ .

**Remark 3.4** *Note that  $\alpha$  and  $\beta$  can be constants or  $o(1)$  as  $N_v \rightarrow \infty$ . For notational simplicity, we omit  $N_v$ .*

**Remark 3.5** *Assumption 3.5 implies  $p = \omega(1/\sqrt{N_v})$ , which is consistent with Assumption 3.4.*

By making assumptions on the underlying rates of error  $\alpha$  and  $\beta$ , we will see that regularity conditions hold for noisy homogeneous and inhomogeneous networks in Appendix 3.5.2 .

### 3.2.2 Biases of standard estimators in noisy networks

Assuming a four-level exposure model and Bernoulli random assignment of treatment, we quantify the biases of standard estimators in homogeneous and inhomogeneous network settings. We begin with the following general result.

**Theorem 3.1** *Assume a four-level exposure model and Bernoulli random assignment of treatment with probability  $p$ . Under Assumptions 3.1 – 3.3, 3.5,  $p = o(1)$ ,  $p = \omega(1/N_v)$  and the potential outcomes are bounded, we have*

$$\begin{aligned} \text{Bias}\left[\tilde{y}_{A\&S}(c_{11})\right] &= -\frac{1}{N_v} \sum_{i=1}^{N_v} \frac{(1-p)^{d_i} [1 - (1-\alpha p)^{N_v-1-d_i}]}{1 - (1-\alpha p)^{N_v-1-d_i} (1 - (1-\beta)p)^{d_i}} \tau_i(c_{11}, c_{10}) + o(1), \\ \text{Bias}\left[\tilde{y}_{A\&S}(c_{10})\right] &= \frac{1}{N_v} \sum_{i=1}^{N_v} [1 - (1-\beta p)^{d_i}] \tau_i(c_{11}, c_{10}), \\ \text{Bias}\left[\tilde{y}_{A\&S}(c_{01})\right] &= -\frac{1}{N_v} \sum_{i=1}^{N_v} \frac{(1-p)^{d_i} [1 - (1-\alpha p)^{N_v-1-d_i}]}{1 - (1-\alpha p)^{N_v-1-d_i} (1 - (1-\beta)p)^{d_i}} \tau_i(c_{01}, c_{00}) + o(1), \\ \text{Bias}\left[\tilde{y}_{A\&S}(c_{00})\right] &= \frac{1}{N_v} \sum_{i=1}^{N_v} [1 - (1-\beta p)^{d_i}] \tau_i(c_{01}, c_{00}), \end{aligned}$$

as  $N_v \rightarrow \infty$ , where  $\tau_i(c_k, c_l) = y_i(c_k) - y_i(c_l)$  and  $d_i$  is the degree of the  $i$ -th vertex in the noise-free network  $G$ .

Theorem 3.1 then directly leads to the following corollary in homogeneous and inhomogeneous network settings.

**Corollary 3.1 (Homogeneous and inhomogeneous)** *Assume a four-level exposure model and Bernoulli random assignment of treatment with  $p$ . In both homogeneous and inhomogeneous network settings, under Assumptions 3.1 – 3.3, 3.5,  $p = o(1)$ ,  $p = \omega(1/N_v)$  and the potential outcomes are bounded, the bias statement in Theorem 3.1 holds.*

The proof of Theorem 3.1 is in supplementary material B.3. Corollary 3.1 directly follows from Theorem 3.1.

Biases of standard estimators in homogeneous and inhomogeneous network settings have the same expressions. Biases of  $\tilde{y}_{A\&S}(c_{11})$  and  $\tilde{y}_{A\&S}(c_{01})$  depend on both  $\alpha$  and  $\beta$ , while biases of  $\tilde{y}_{A\&S}(c_{10})$  and  $\tilde{y}_{A\&S}(c_{00})$  only depend on  $\beta$ . Biases of  $\tilde{y}_{A\&S}(c_{11})$  and  $\tilde{y}_{A\&S}(c_{10})$  are related to  $\tau(c_{11}, c_{10})$ . And biases of  $\tilde{y}_{A\&S}(c_{01})$  and  $\tilde{y}_{A\&S}(c_{00})$  are related to  $\tau(c_{01}, c_{00})$ . These relationships follow because the network noise affects observed edges but not treatment status.

Let  $\tilde{y}_{A\&S,i}(c_k)$  denote the Aronow and Samii estimator for  $y_i(c_k)$  in noisy networks, which corresponds to the  $i$ -th element of  $\tilde{y}_{A\&S}(c_k)$  in (3.6). We summarize in the following table the asymptotic biases of  $\tilde{y}_{A\&S,i}(c_k)$  for high (top row) and low (bottom row) degree nodes.

**Table 3.1:** The asymptotic biases of  $\tilde{y}_{A\&S,i}(c_k)$  for high (top row) and low (bottom row) degree nodes.

	Bias $[\tilde{y}_{A\&S,i}(c_{11})]$	Bias $[\tilde{y}_{A\&S,i}(c_{10})]$	Bias $[\tilde{y}_{A\&S,i}(c_{01})]$	Bias $[\tilde{y}_{A\&S,i}(c_{00})]$
$d_i = \omega(1/p)$	$o(1)$	$\tau_i(c_{11}, c_{10})$	$o(1)$	$\tau_i(c_{01}, c_{00})$
$d_i = o(1/p)$	$-\tau_i(c_{11}, c_{10})$	$o(1)$	$-\tau_i(c_{01}, c_{00})$	$o(1)$

We see that there are four cases where  $\tilde{y}_{A\&S,i}(c_k)$  is asymptotically unbiased. The reason is that the corresponding entries in the expected confusion matrix (3.5) go to 0. For the other four cases, the corresponding entries in the expected confusion matrix approach 1, which leads to nontrivial biases. Note that the asymptotic biases of  $\tilde{y}_{A\&S,i}(c_k)$  is between 0 and the corresponding  $\pm\tau_i(c_k, c_l)$  when  $d_i = \Theta(1/p)$ .

### 3.2.3 Variances of standard estimators in noisy networks

We analyze the variances of standard estimators in homogeneous and inhomogeneous network settings.

**Theorem 3.2 (Homogeneous)** *Assume a four-level exposure model and Bernoulli random assignment of treatment with probability  $p$ . In the homogeneous network*

setting, under Assumptions 3.1 - 3.5, for all  $c_k$ , we have  $\text{Var}[\tilde{y}_{A\&S}(c_k)] = o(1)$  as  $N_v \rightarrow \infty$ .

**Theorem 3.3 (Inhomogeneous)** *Assume a four-level exposure model and Bernoulli random assignment of treatment with probability  $p$ . In the inhomogeneous network setting, under Assumptions 3.1 - 3.5,  $\lambda = \Theta(p)$  and  $\lambda > p$ , we have  $\text{Var}[\tilde{y}_{A\&S}(c_k)] = o(1)$  for all  $c_k$  as  $N_v \rightarrow \infty$ .*

Note that the variances go to zero as the number of nodes tends towards infinity for both cases. Therefore, in noisy networks, the bias would appear to be the primary concern for estimating average causal effects.

### 3.3 Accounting for network noise

As we saw in Section 3.2, standard estimators are biased in both homogeneous and inhomogeneous network settings. Thus, it is important to have new estimators for bias reduction. We present method-of-moments estimators in Section 3.3.1, and show unbiasedness and consistency under a four-level exposure model and Bernoulli random assignment of treatment in Section 3.3.2. The method-of-moments estimators require either knowledge or consistent estimators of  $\alpha$  and  $\beta$ . We adopt the estimators in Chang et al. (2020), which require at least three replicates of the observed network.

#### 3.3.1 Method-of-moments estimators

We construct method-of-moments estimators (MME) by reweighting the observed outcomes based on the expected confusion matrix. For convenience, we denote

$$\begin{aligned} \mathbf{y}_i &= [y_i(c_{11}), y_i(c_{10}), y_i(c_{01}), y_i(c_{00})]^\top, \\ \mathbb{1}(\mathbf{x}_i) &= [I_{\{f(\mathbf{Z}, \mathbf{x}_i)=c_{11}\}}, I_{\{f(\mathbf{Z}, \mathbf{x}_i)=c_{10}\}}, I_{\{f(\mathbf{Z}, \mathbf{x}_i)=c_{01}\}}, I_{\{f(\mathbf{Z}, \mathbf{x}_i)=c_{00}\}}]^\top, \\ \mathbb{1}(\tilde{\mathbf{X}}_i) &= [I_{\{f(\mathbf{Z}, \tilde{\mathbf{X}}_i)=c_{11}\}}, I_{\{f(\mathbf{Z}, \tilde{\mathbf{X}}_i)=c_{10}\}}, I_{\{f(\mathbf{Z}, \tilde{\mathbf{X}}_i)=c_{01}\}}, I_{\{f(\mathbf{Z}, \tilde{\mathbf{X}}_i)=c_{00}\}}]^\top. \end{aligned}$$

We then combine the observed outcome  $\mathbb{1}(\mathbf{x}_i)^\top \mathbf{y}_i$  and the observed exposure level into a vector, denoted by  $\tilde{\mathbf{y}}_i$ ,

$$\tilde{\mathbf{y}}_i = \mathbb{1}(\tilde{\mathbf{X}}_i) \cdot \mathbb{1}(\mathbf{x}_i)^\top \mathbf{y}_i. \quad (3.7)$$

By taking the expectation with respect to treatment and network noise, we obtain

$$\mathbb{E}[\tilde{\mathbf{y}}_i] = \mathbf{P}_i \cdot \mathbf{y}_i, \quad (3.8)$$

Note that  $\mathbf{P}_i$  depends on  $d_i$ ,  $\alpha$  and  $\beta$ . Therefore, we use  $\mathbf{P}(d_i, \alpha, \beta)$  for explicitness.

The method of moments estimator for  $\mathbf{y}_i$  is defined as

$$\tilde{\mathbf{y}}_{\text{MME},i} = \mathbf{P}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \tilde{\mathbf{y}}_i, \quad (3.9)$$

where

$$\hat{d}_i = \frac{\tilde{d}_i - (N_v - 1)\hat{\alpha}}{1 - \hat{\alpha} - \hat{\beta}}. \quad (3.10)$$

The values  $\hat{\alpha}$  and  $\hat{\beta}$  are consistent estimators of  $\alpha$  and  $\beta$ , which we provide an example of consistent estimators later. If  $\alpha$  and  $\beta$  are known, we substitute  $\hat{\alpha}$  and  $\hat{\beta}$  in (3.9) and (3.10) with the known values, and this does not change the asymptotic behavior we state in Section 3.3.2.

We define the method-of-moments estimator for  $\sum_{i=1}^{N_v} \mathbf{y}_i / N_v$

$$\tilde{\mathbf{y}}_{\text{MME}} = \frac{1}{N_v} \sum_{i=1}^{N_v} \left\{ \tilde{\mathbf{y}}_{\text{MME},i} \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} + \tilde{\mathbf{y}}_{\text{A\&S},i} \cdot I_{\{\hat{d}_i = o(1/p) \text{ or } \omega(1/p)\}} \right\}, \quad (3.11)$$

where  $\tilde{\mathbf{y}}_{\text{A\&S},i}$  is the Aronow and Samii estimator of node  $i$  in the noisy network. Recall from the bias statements in Table 3.1 that  $\tilde{y}_{\text{A\&S},i}(c_{11})$  and  $\tilde{y}_{\text{A\&S},i}(c_{01})$  are asymptotically unbiased for nodes with high degrees. And  $\tilde{y}_{\text{A\&S},i}(c_{10})$  and  $\tilde{y}_{\text{A\&S},i}(c_{00})$  are asymptotically unbiased for small degree nodes. Therefore, we do not need to correct biases for those cases. We will show that  $\tilde{\mathbf{y}}_{\text{MME},i}$  is asymptotically unbiased with small variance for nodes with degree on the order of  $1/p$  in Theorem 3.4 and

3.5. Otherwise, asymptotically unbiased estimators with small variances may not exist due to the structure of this specific four-level exposure model. As we saw,  $\mathbb{E}[\tilde{y}_{A\&S,i}(c_{11})] \rightarrow y_i(c_{10})$  and  $\mathbb{E}[\tilde{y}_{A\&S,i}(c_{01})] \rightarrow y_i(c_{00})$  for small degree nodes, while  $\mathbb{E}[\tilde{y}_{A\&S,i}(c_{10})] \rightarrow y_i(c_{11})$  and  $\mathbb{E}[\tilde{y}_{A\&S,i}(c_{00})] \rightarrow y_i(c_{01})$  for high degree nodes. These mean that we lose almost all information about  $y_i(c_{11})$  and  $y_i(c_{01})$  for small degree nodes, and  $y_i(c_{10})$  and  $y_i(c_{00})$  for high degree nodes.

In general, we suggest to use terms of the same orders of magnitude in (3.11) to approximate  $\Theta(\cdot)$ . That is, writing  $1/p = a \times 10^b$ , where  $1/\sqrt{10} \leq a < \sqrt{10}$ , we represent the order of magnitude with  $b$ . Next, we rewrite  $\tilde{\mathbf{y}}_{\text{MME}}$  as

$$\tilde{\mathbf{y}}_{\text{MME}} = \frac{1}{N_v} \sum_{i=1}^{N_v} \left\{ \tilde{\mathbf{y}}_{\text{MME},i} \cdot I_{\{C_1 \leq \hat{d}_i < C_2\}} + \tilde{\mathbf{y}}_{A\&S,i} \cdot I_{\{\hat{d}_i < C_1 \text{ or } \hat{d}_i \geq C_2\}} \right\}, \quad (3.12)$$

where  $C_1 = 10^b/\sqrt{10}$  and  $C_2 = \sqrt{10} \cdot 10^b$ . For sparse and inhomogeneous networks with small sample sizes,  $C_1$  may be close to the average degree and thus we recommend to compute

$$\tilde{\mathbf{y}}_{\text{MME}} = \frac{1}{N_v} \sum_{i=1}^{N_v} \left\{ \tilde{\mathbf{y}}_{\text{MME},i} \cdot I_{\{\hat{d}_i \geq 1\}} + \tilde{\mathbf{y}}_{A\&S,i} \cdot I_{\{\hat{d}_i < 1\}} \right\}. \quad (3.13)$$

**Remark 3.6** *As we will see later, in this specific four-level exposure model,  $\tilde{\mathbf{y}}_{\text{MME}}$  is asymptotically unbiased and consistent in both homogeneous and inhomogeneous network settings.*

Our estimators require knowledge or, more realistically, consistent estimates of the parameters  $\alpha$  and  $\beta$  governing the noise. Here, we adopt the consistent MME estimators in Chang et al. (2020), which require at least three replicates of the observed network. Define relevant quantities as follows:

$$\begin{aligned} u_1 &= (1 - \delta)\alpha + \delta(1 - \beta), \\ u_2 &= (1 - \delta)\alpha(1 - \alpha) + \delta\beta(1 - \beta), \\ u_3 &= (1 - \delta)\alpha(1 - \alpha)^2 + \delta\beta^2(1 - \beta), \end{aligned}$$

where  $\delta$  is the edge density in the true network  $G$ ,  $u_1$  is the expected edge density in one observed network,  $u_2$  is the expected density of edge differences in two observed networks, and  $u_3$  is the average probability of having an edge between two arbitrary nodes in one observed network but no edge between the same nodes in the other two observed networks. The method-of-moments estimators for  $u_1$ ,  $u_2$  and  $u_3$  are

$$\begin{aligned}\hat{u}_1 &= \frac{2}{N_v(N_v - 1)} \sum_{i < j} \tilde{A}_{i,j}, \\ \hat{u}_2 &= \frac{1}{N_v(N_v - 1)} \sum_{i < j} |\tilde{A}_{i,j,*} - \tilde{A}_{i,j}|, \\ \hat{u}_3 &= \frac{2}{3N_v(N_v - 1)} \sum_{i < j} I(\text{Exactly one of } \tilde{A}_{i,j,**}, \tilde{A}_{i,j,*}, \tilde{A}_{i,j} \text{ equals 1}),\end{aligned}\tag{3.14}$$

where  $\tilde{\mathbf{A}}_* = (\tilde{A}_{i,j,*})_{N_v \times N_v}$ ,  $\tilde{\mathbf{A}}_{**} = (\tilde{A}_{i,j,**})_{N_v \times N_v}$  are independent and identically distributed replicates of  $\tilde{\mathbf{A}}$ . Calculation of the estimators  $\hat{\alpha}$  and  $\hat{\beta}$  can be accomplished as detailed in Algorithm 3.1 below.

---

**Algorithm 3.1** Consistent estimators  $\hat{\alpha}$  and  $\hat{\beta}$

---

**Input:**  $\tilde{\mathbf{A}} = (\tilde{A}_{i,j})_{N_v \times N_v}$ ,  $\tilde{\mathbf{A}}_* = (\tilde{A}_{i,j,*})_{N_v \times N_v}$ ,  $\tilde{\mathbf{A}}_{**} = (\tilde{A}_{i,j,**})_{N_v \times N_v}$ ,  $\alpha_0$ ,  $\varepsilon$

**Output:**  $\hat{\alpha}$ ,  $\hat{\beta}$

Compute  $\hat{u}_1$ ,  $\hat{u}_2$ ,  $\hat{u}_3$  defined in (3.14);

Initialize  $\hat{\alpha} = \alpha_0$ ,  $\alpha_0 = \hat{\alpha} + 10\varepsilon$ ;

**while**  $|\hat{\alpha} - \alpha_0| > \varepsilon$  **do**

$$\alpha_0 \leftarrow \hat{\alpha}, \quad \hat{\beta} \leftarrow \frac{\hat{u}_2 - \alpha_0 + \hat{u}_1 \alpha_0}{\hat{u}_1 - \alpha_0}, \quad \hat{\delta} \leftarrow \frac{(\hat{u}_1 - \alpha_0)^2}{\hat{u}_1 - \hat{u}_2 - 2\hat{u}_1 \alpha_0 + \alpha_0^2}, \quad \hat{\alpha} \leftarrow \frac{\hat{u}_3 - \hat{\delta} \hat{\beta}^2 (1 - \hat{\beta})}{(1 - \hat{\delta})(1 - \alpha_0)^2}.$$


---

### 3.3.2 Asymptotic unbiasedness and consistency

We consider the asymptotic behavior of the method-of-moments estimators  $\tilde{y}_{\text{MME}}(c_k)$  as  $N_v \rightarrow \infty$ .

**Theorem 3.4 (Homogeneous)** *Assume a four-level exposure model and Bernoulli random assignment of treatment with  $p$ . In the homogeneous network setting, under Assumptions 3.1 - 3.5,  $\tilde{y}_{\text{MME}}(c_k)$  is an asymptotically unbiased and consistent estimator of  $\bar{y}(c_k)$  for all  $c_k$ .*

**Theorem 3.5 (Inhomogeneous)** *Assume a four-level exposure model and Bernoulli random assignment of treatment with  $p$ . In the inhomogeneous network setting, under Assumptions 3.1 - 3.5,  $\lambda = \Theta(p)$  and  $\lambda > p$ ,  $\tilde{y}_{MME}(c_k)$  is an asymptotically unbiased and consistent estimator of  $\bar{y}(c_k)$  for all  $c_k$ .*

Note that  $\tilde{y}_{MME}(c_k)$  is an asymptotically unbiased and consistent estimator of  $\bar{y}(c_k)$  in both homogeneous and inhomogeneous network settings. Proofs of Theorem 3.4 and 3.5 appear in supplementary material B.3.

### 3.4 Numerical illustration: British secondary school contact networks

We conduct some simulations to illustrate the finite sample properties of the proposed estimation methods. We consider the data and network construction described in Kucharski et al. (2018). These data were collected from 460 unique participants across four rounds of data collection conducted between January and June 2015 in year 7 groups in four UK secondary schools, with 7,315 identifiable contacts reported in total. They used a process of peer nomination as a method for data collection: students were asked, via the research questionnaire, to list the six other students in year 7 at their school that they spend the most time with. For each pair of participants in a specific round of data collection, a single link was defined if either one of the participants reported a contact between the pair (i.e. there was at least one unidirectional link, in either direction). Our analysis focuses on the single link contact network.

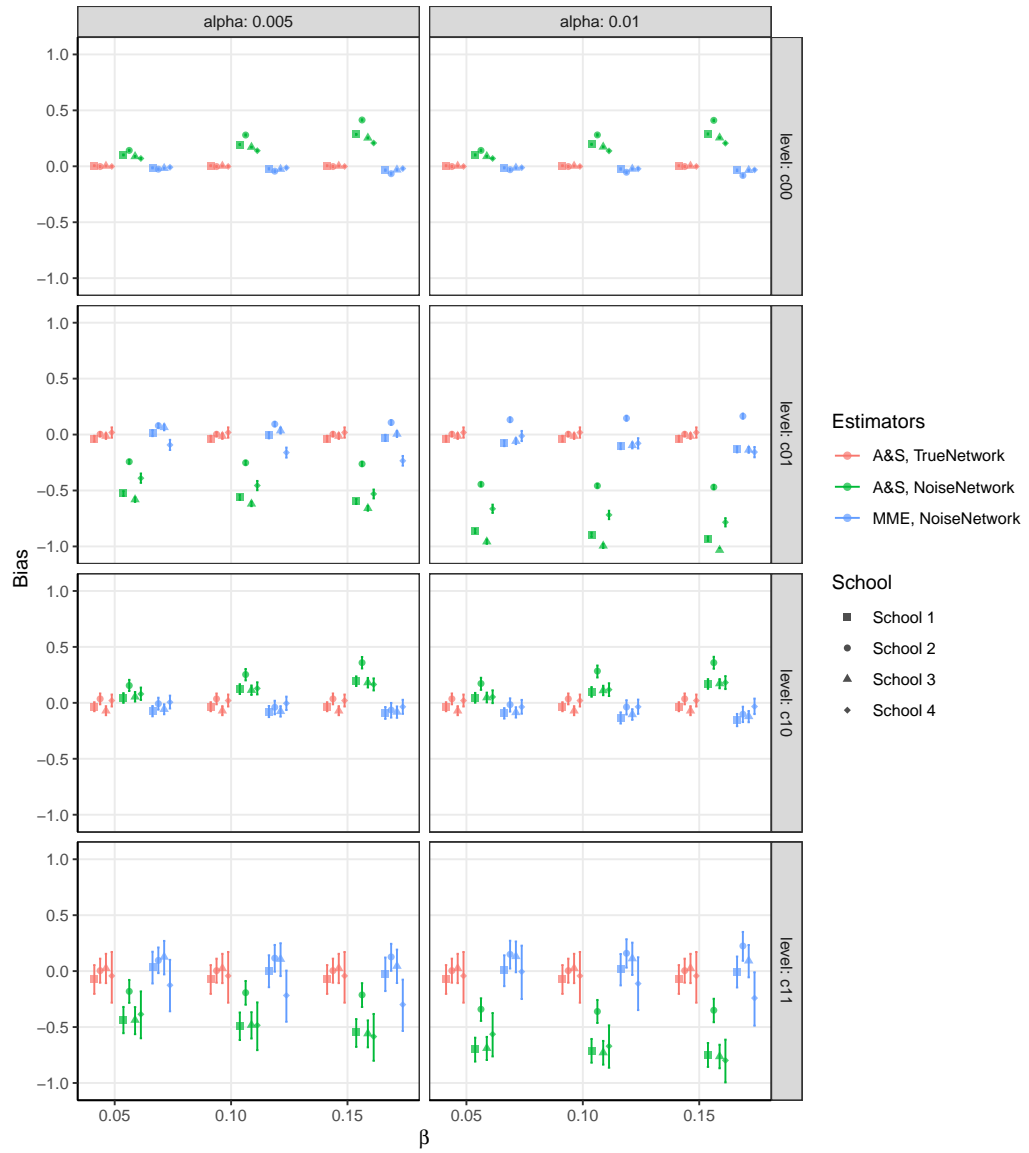
For each school, we construct a ‘true’ adjacency matrix  $\mathbf{A}$ : if an edge occurs between a pair of vertices more than once in four rounds, we view that pair to have a true edge. The noisy, observed adjacency matrices  $\tilde{\mathbf{A}}$ ,  $\tilde{\mathbf{A}}_*$ ,  $\tilde{\mathbf{A}}_{**}$  are generated according to (3.1). We set  $\alpha = 0.005$  or  $0.010$ , and  $\beta = 0.05$ ,  $0.10$ , or  $0.15$ . We assume that both  $\alpha$  and  $\beta$  are unknown. For treatment effects we adopt a simple model in the

spirit of the ‘diluted effects’ model of Rosenbaum (Rosenbaum (1999)) and suppose  $y_i(c_{11}) = 10$ ,  $y_i(c_{10}) = 7$ ,  $y_i(c_{01}) = 5$ ,  $y_i(c_{00}) = 1$ . We set  $p = 0.1$  and explore the effect of  $\alpha$ ,  $\beta$  on the performance of estimators  $\tilde{y}(\cdot)$ .

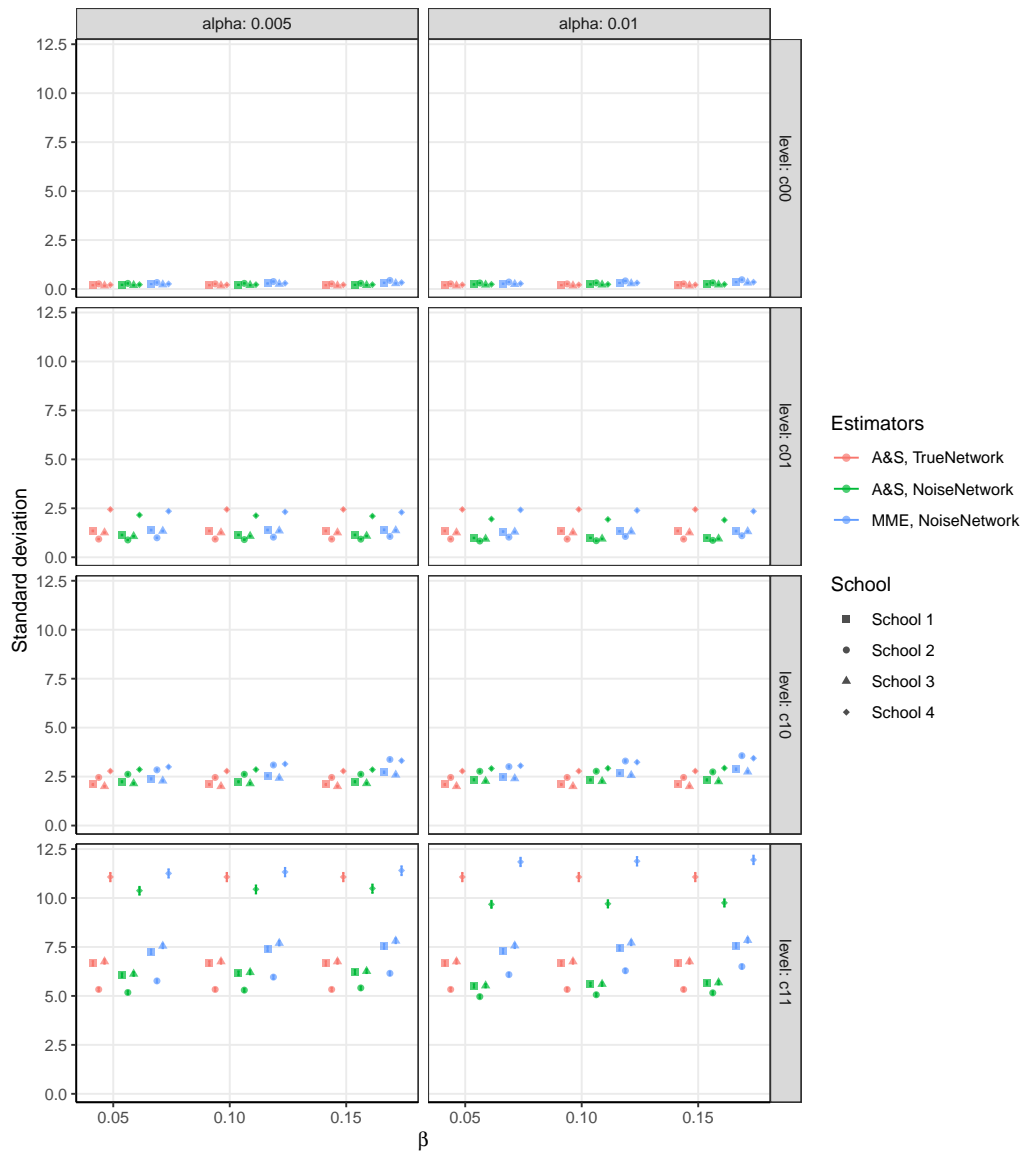
We run Monte Carlo simulation of 10,000 trials and compute three kinds of estimators: Aronow and Samii estimators in noise-free networks, Aronow and Samii estimators in noisy networks, and method-of-moments estimators in noisy networks. For the method-of-moments estimators, we first obtain estimators  $\hat{\alpha}$  and  $\hat{\beta}$  by Algorithm 3.1. The networks are sparse with inhomogeneous degree distribution and small sizes, so we compute  $\tilde{\mathbf{y}}_{\text{MME}}$  by (3.13). Also, we run 1,000 times bootstrap resampling of estimators to obtain 95% confidence intervals for biases and standard deviations of estimators. Biases and standard deviations are shown in Figure 3.1 and 3.2. Error bars are 95% confidence intervals.

From the plots, we see that method-of-moments estimators outperform Aronow and Samii estimators in noisy networks, and essentially perform the same on noisy networks as Aronow and Samii estimators do on noise-free networks (same zero biases with, at times, just slightly larger standard deviations). Aronow and Samii estimators in noisy networks underestimate  $\bar{y}(c_{11})$  and  $\bar{y}(c_{01})$  and overestimate  $\bar{y}(c_{10})$  and  $\bar{y}(c_{00})$ . And the biases of Aronow and Samii estimators for  $\bar{y}(c_{11})$  and  $\bar{y}(c_{01})$  increase as  $\alpha$  and  $\beta$  increase, while the biases of Aronow and Samii estimators for  $\bar{y}(c_{01})$  and  $\bar{y}(c_{00})$  only depend on  $\beta$ . The biases of method-of-moments estimators are close to zero in all cases. In addition, standard deviations of the three types estimators are similar in all cases. The standard deviations of estimators in School 4 are larger than those in other schools because the network size in School 4 is relatively small.

**Figure 3.1:** Biases of estimators for  $\bar{y}(\cdot)$  for noisy networks in four schools. Error bars are 95% confidence intervals.



**Figure 3.2:** Standard deviations of estimators for  $\bar{y}(\cdot)$  for noisy networks in four schools. Error bars are 95% confidence intervals.



## Data and code accessibility

No primary data are used in this chapter. Secondary data source is taken from Kucharski et al. (2018). These data and the code necessary to reproduce the results in this chapter are available at <https://github.com/KolaczykResearch/CausInfNoisyNet>.

## 3.5 Appendix

In this appendix, we provide arguments for the consistency of contrast estimates in noise-free networks and regularity conditions in noisy networks. And we present the exposure probabilities in the generalized four-level exposure model.

### 3.5.1 Consistency of contrast estimates in noise-free networks

We first establish conditions for the estimator  $\hat{\tau}(c_k, c_l)$  to converge to  $\tau(c_k, c_l)$  as  $N_v \rightarrow \infty$ . We will show that, under two regularity conditions,  $\hat{\tau}(c_k, c_l) \xrightarrow{P} \tau(c_k, c_l)$  as  $N_v \rightarrow \infty$ . Note that these conditions are similar to but slightly more general than the conditions in Aronow and Samii (2017).

**Condition 3.1** *For all values  $i$  and  $c_k$ ,  $|y_i(c_k)| \leq c < \infty$ ,  $p_i^e(c_k) > 0$  and  $\sum_{i=1}^{N_v} 1/p_i^e(c_k) = o(N_v^2)$ , where  $c$  is a constant.*

We will also make an assumption about the amount of dependence among exposure conditions in the population. Let  $p_{ij}^e(c_k) = \sum_{\mathbf{z}} p_{\mathbf{z}} I_{\{f(\mathbf{z}, \mathbf{x}_i)=c_k\}} I_{\{f(\mathbf{z}, \mathbf{x}_j)=c_k\}}$ .

**Condition 3.2** *For all values  $c_k$ ,  $\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} |p_{ij}^e(c_k)/(p_i^e(c_k)p_j^e(c_k)) - 1| = o(N_v^2)$ .*

Condition 3.2 implies that the amount of pairwise clustering in exposure conditions is limited in scope as  $N_v$  grows. Condition 3.2 can be relaxed, though Condition 3.1 would likely need to be strengthened accordingly.

**Proposition 3.1** *Given Conditions 3.1 and 3.2,  $\hat{\tau}(c_k, c_l) \xrightarrow{P} \tau(c_k, c_l)$  as  $N_v \rightarrow \infty$ .*

Assuming the four-level exposure model in (3.2) and Bernoulli random assignment of treatment with probability  $p$ , we consider the consistency of the estimator  $\hat{\tau}(c_k, c_l)$  in two typical classes of networks: homogeneous and inhomogeneous.

**Proposition 3.2 (Homogeneous)** *Assume a four-level exposure model and Bernoulli random assignment of treatment with  $p$ . In the homogeneous network setting, under Assumption 3.4,  $\hat{\tau}(c_k, c_l) \xrightarrow{P} \tau(c_k, c_l)$  as  $N_v \rightarrow \infty$ .*

**Proposition 3.3 (Inhomogeneous)** *Assume a four-level exposure model and Bernoulli random assignment of treatment with  $p$ . In the inhomogeneous network setting, under Assumption 3.4,  $\lambda = \Theta(p)$  and  $\lambda > p$ , we have  $\hat{\tau}(c_k, c_l) \xrightarrow{P} \tau(c_k, c_l)$  as  $N_v \rightarrow \infty$ .*

Proofs for Propositions 3.1 – 3.3 appear in the supplementary material B.1.

Note that, under Assumption 3.4, Condition 3.1 does not hold for levels  $c_{10}$  and  $c_{00}$  when the degrees follow a Pareto distribution with shape  $\zeta > 1$ . This is because there are more high degree nodes, and  $1/p_i^e(c_{10})$  and  $1/p_i^e(c_{00})$  increase exponentially when the degree  $d_i$  increases. See supplementary material B.5 for the proof.

### 3.5.2 Standard estimators in noisy networks

Recall that under the Condition 3.1 and 3.2,  $\hat{\tau}(c_k, c_l) \xrightarrow{P} \tau(c_k, c_l)$  as  $N_v \rightarrow \infty$ . By making assumptions on underlying rates of error  $\alpha$  and  $\beta$ , we will show that similar regularity conditions hold for noisy homogeneous and inhomogeneous networks. These conditions will then be used in our characterization of bias and variance in Sections 3.2.2 and 3.2.3. Define  $\tilde{p}_{ij}^e(c_k) = \sum_{\mathbf{z}} p_{\mathbf{z}} I_{\{f(\mathbf{z}, \tilde{\mathbf{x}}_i) = c_k\}} I_{\{f(\mathbf{z}, \tilde{\mathbf{x}}_j) = c_k\}}$ .

**Proposition 3.4 (Homogeneous)** *Assume a four-level exposure model and Bernoulli random assignment of treatment with  $p$ . In the homogeneous network setting, under Assumptions 3.1 - 3.5, for all values  $i$  and  $c_k$ ,  $\mathbb{P}(\tilde{p}_i^e(c_k) > 0) \rightarrow 1$ ,  $\mathbb{E}[\sum_{i=1}^{N_v} I_{\{\tilde{p}_i^e(c_k) > 0\}} / \tilde{p}_i^e(c_k)] = o(N_v^2)$ , and  $\mathbb{E}[\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} I_{\{\tilde{p}_i^e(c_k) > 0\}} I_{\{\tilde{p}_j^e(c_k) > 0\}} | \tilde{p}_{ij}^e(c_k) / (\tilde{p}_i^e(c_k) \tilde{p}_j^e(c_k)) - 1 |] = o(N_v^2)$ .*

**Proposition 3.5 (Inhomogeneous)** *Assume a four-level exposure model and Bernoulli random assignment of treatment with  $p$ . In the inhomogeneous network setting, under Assumptions 3.1- 3.5,  $\lambda = \Theta(p)$  and  $\lambda > p$ , the statements in Proposition 3.4 hold for all values  $i$  and  $c_k$ .*

See supplementary material B.2 for proofs of Propositions 3.4 and 3.5.

### 3.5.3 The exposure probabilities in the generalized four-level exposure model

In the generalized four-level exposure model (5.1) and Bernoulli random assignment of treatment with  $p$ , for each individual  $i$ , the four exposure probabilities are as follows.

$$\begin{aligned}
 p_i^e(c_{11'}) &= p \sum_{x=m_i}^{d_i} \binom{d_i}{x} p^x (1-p)^{d_i-x}, \\
 p_i^e(c_{10'}) &= p \sum_{x=0}^{(m_i-1) \wedge d_i} \binom{d_i}{x} p^x (1-p)^{d_i-x}, \\
 p_i^e(c_{01'}) &= (1-p) \sum_{x=m_i}^{d_i} \binom{d_i}{x} p^x (1-p)^{d_i-x}, \\
 p_i^e(c_{00'}) &= (1-p) \sum_{x=0}^{(m_i-1) \wedge d_i} \binom{d_i}{x} p^x (1-p)^{d_i-x}.
 \end{aligned} \tag{3.15}$$

## Chapter 4

# Estimation of local time-varying reproduction numbers in noisy surveillance data

In the context of an epidemic, institutions and similar entities can be expected to adopt various intervention measures. Distinguishing between cases infected with the disease due to local transmission, versus due to importation, thus becomes important for understanding the effectiveness of these interventions. Realistically, we can expect identification of cases as local or imported to be imperfect. We study the propagation of such errors on local time-varying reproduction number. In addition, we propose a Bayesian framework for estimation of the true local time-varying reproduction number when identification errors exist. And we illustrate the practical performance of our estimator through simulation studies and with outbreaks of COVID-19 in Hong Kong and Victoria, Australia.

The organization of this chapter is as follows. In Section 4.1 we introduce the problem and review related literature. In Section 4.2 we show the bias of the noisy local time-varying reproduction number (i.e., a standard estimate that assumes perfect identification, when instead they are noisy), and propose a Bayesian hierarchical framework to estimate the true local time-varying reproduction number with imperfect knowledge. Section 4.3 reports the practical performance of our estimators through simulation studies and with SARS-CoV-2 infections in Hong Kong and Australia.

## 4.1 Introduction

The local time-varying reproduction number,  $R_*^{\text{local}}(t)$ , is an important quantity to monitor the infectiousness and transmissibility of diseases and, therefore, to design and adjust public health responses during an outbreak. Recent examples include monitoring transmission of the COVID-19 pandemic and demonstrating the efficacy of non-pharmaceutical interventions in more than 100 countries (You et al. (2020); Li et al. (2020b); Rubin et al. (2020); Abbott et al. (2020)). The value of  $R_*^{\text{local}}(t)$  represents the expected number of secondary local cases arising from a primary case infected at time  $t$ . Different formal definitions of  $R_*^{\text{local}}(t)$  have been proposed, and a number of methods are available to estimate this quantity. The widely used is an estimator of the instantaneous reproduction number that is defined as the ratio of the expected number of incident locally infected cases at time  $t$  to the expected total infectiousness of infected individuals at time  $t$  (Thompson et al. (2019); Cori et al. (2013)).

Distinguishing local cases from imported cases is essential to estimation of the local time-varying reproduction number. However, surveillance data generally is available only up to some level of error. For example, if we are unable to identify the correct source of infection from contact tracing or genetic information, imported cases might be misclassified as local cases, and vice versa. Such misclassification error is recognized as one limitation of estimating  $R_*^{\text{local}}(t)$  in the COVID-19 outbreak (Chong et al. (2020); Arroyo Marioli et al. (2020)). We investigate how identification error impacts on the estimation of the instantaneous reproduction number and, thus, on our understanding of diseases transmission dynamics.

Extensive work regarding improving inference of time-varying reproduction numbers has been done. For instance, there have been efforts to estimate the serial interval that is used to compute the total infectiousness for  $R_*^{\text{local}}(t)$  estimation, including

Bayesian parametric estimation using data augmentation Markov Chain Monte Carlo (Reich et al. (2009)), and a cure model for limited follow-up data (Ma et al. (2020)). Many studies have explored the effects of imperfect detection and estimated the true infection prevalence (Miller et al. (2012); McClintock et al. (2010); Cui et al. (2013); Arroyo Marioli et al. (2020)). But, to our best knowledge, there has been little attention to date given towards accounting for identification errors of local and imported cases.

Our contribution is to quantify how such errors propagate to the local time-varying reproduction number, and to provide estimators for  $R_*^{\text{local}}(t)$  when contact tracing survey information is available. Adopting the definition of  $R_*^{\text{local}}(t)$  proposed by Thompson et al. (2019), we characterize the impact of identification errors on the bias of noisy local time-varying reproduction numbers. Our work shows that, in general, the bias can be expected to be nontrivial. Accordingly, we propose a Bayesian framework to estimate the true local time-varying reproduction number. Numerical simulation suggests that high accuracy is possible for estimating local time-varying reproduction numbers in outbreaks of even modest size. We illustrate the practical use of our estimators in the context of COVID-19 pandemic in Hong Kong and Victoria, Australia.

## 4.2 Methods

In this section, we first quantify the bias of the noisy local time-varying reproduction number when misidentification occurs in the surveillance data. We then build a Bayesian hierarchical framework to estimate true local time-varying reproduction numbers. We also propose a method to estimate misidentification rates based on contact tracing survey data, which informs the prior distribution in the model.

### 4.2.1 Notation

We provide essential notation and background here. The number of newly infected cases at time  $t$ ,  $I_*(t)$ , is the sum of the numbers of local ( $I_*^{\text{local}}(t)$ ) and imported ( $I_*^{\text{imported}}(t)$ ) cases. If one assumes independence between calendar time and the generation interval (the duration between the infection time of a secondary infectee and the infection time of its infector),  $g(s)$ , then the local time-varying reproduction number is defined as (Thompson et al. (2019))

$$R_*^{\text{local}}(t) = \frac{\mu_*^{\text{local}}(t)}{\int_0^\infty g(s)\mu_*(t-s)ds}, \quad (4.1)$$

where  $\mu_*^{\text{local}}(t) = \mathbb{E}[I_*^{\text{local}}(t)]$  and  $\mu_*(t) = \mathbb{E}[I_*(t)]$ .

In reality, we only know the serial interval and the number of diagnosed cases. Let  $I(t)$ ,  $I^{\text{local}}(t)$  and  $I^{\text{imported}}(t)$  be the numbers of total diagnosed cases, local diagnosed cases, and imported diagnosed cases at time  $t$ , respectively. Then, we define a realistic local time-varying reproduction number as

$$R^{\text{local}}(t) = \frac{\mu^{\text{local}}(t)}{\int_0^\infty w(s)\mu(t-s)ds}, \quad (4.2)$$

where  $w(s)$  is the serial interval (the time between the start of symptoms in the infector and onset of symptoms in the infectee),  $\mu^{\text{local}}(t) = \mathbb{E}[I^{\text{local}}(t)]$  and  $\mu(t) = \mathbb{E}[I(t)]$ . Note that the serial interval corresponds to date of symptom onset. One can estimate symptom onset dates by back calculation of report dates (Li and White (2020)).

Realistically, we can expect identification of cases as local or imported to be imperfect. Let  $\tilde{I}^{\text{local}}(t)$  and  $\tilde{I}^{\text{imported}}(t)$  be the number of new local and imported cases reported at time  $t$ , with identification error. Thus, we define a noisy local time-varying reproduction number as

$$\tilde{R}^{\text{local}}(t) = \frac{\tilde{\mu}^{\text{local}}(t)}{\int_0^\infty w(s)\mu(t-s)ds}, \quad (4.3)$$

where  $\tilde{\mu}^{\text{local}}(t) = \mathbb{E}[\tilde{I}^{\text{local}}(t)]$ . The definition of  $\tilde{R}^{\text{local}}(t)$  in (4.3) comes from an argument that mimics the original argument using Poisson arrivals in Fraser (2007). Specifically, we suppose that we observe a Poisson stream  $\tilde{I}^{\text{local}}(t)$  that is a function of calendar time  $t$  in terms of the transmissibility, denoted  $\tilde{\beta}^{\text{local}}(t, s)$ , an arbitrary function of calendar time  $t$  and time since infection  $s$ . Then,  $\tilde{\mu}^{\text{local}}(t)$  follows the so-called renewal equation

$$\tilde{\mu}^{\text{local}}(t) = \int_0^{\infty} \tilde{\beta}^{\text{local}}(t, s) \mu(t-s) ds. \quad (4.4)$$

Following Fraser (2007), we have

$$\tilde{\beta}^{\text{local}}(t, s) = \tilde{R}^{\text{local}}(t) w(s). \quad (4.5)$$

Inserting (4.5) into (4.4) yields the definition of  $\tilde{R}^{\text{local}}(t)$  in (4.3).

Our interest is in characterizing the manner in which the uncertainty in  $\tilde{I}^{\text{local}}(t)$  and  $\tilde{I}^{\text{imported}}(t)$  propagates to the local time-varying reproduction number, and providing estimators of  $R^{\text{local}}(t)$  to account for identification errors.

#### 4.2.2 Bias of the noisy local time-varying reproduction number

We quantify the bias of the noisy local time-varying reproduction number in (4.3) when misidentification occurs. We begin by defining a model for  $\tilde{I}^{\text{local}}(t)$  and  $\tilde{I}^{\text{imported}}(t)$ . Let  $\alpha_0$  denote the probability that an imported case is misidentified as local, and  $\alpha_1$  the probability that a local case is misidentified as imported. Then, a simple model is

$$\begin{aligned} \tilde{I}^{\text{local}}(t) | I^{\text{local}}(t), I^{\text{imported}}(t), \alpha_0, \alpha_1 &\sim \text{Bin}(I^{\text{local}}(t), 1 - \alpha_1) + \text{Bin}(I^{\text{imported}}(t), \alpha_0), \\ \tilde{I}^{\text{imported}}(t) &= I^{\text{local}}(t) + I^{\text{imported}}(t) - \tilde{I}^{\text{local}}(t). \end{aligned} \quad (4.6)$$

Under independence, the first relationship in (4.6) is directly obtained by the definition of  $\alpha_0$  and  $\alpha_1$ . And the second equation in (4.6) is due to the fact that the total number of cases reported at time  $t$  is not affected by the misidentification.

By (4.6), the relationship between  $\tilde{\mu}^{\text{local}}(t)$  and  $\mu^{\text{local}}(t)$  is

$$\tilde{\mu}^{\text{local}}(t) = (1 - \alpha_1)\mu^{\text{local}}(t) + \alpha_0\mu^{\text{imported}}(t), \quad (4.7)$$

where  $\mu^{\text{imported}}(t) = \mathbb{E}(I^{\text{imported}}(t))$ . Direct computation yields

$$\tilde{R}^{\text{local}}(t) = \left(1 - \alpha_1 + \alpha_0 \frac{\mu^{\text{imported}}(t)}{\mu^{\text{local}}(t)}\right) R^{\text{local}}(t) \quad (4.8)$$

when  $\mu^{\text{local}}(t) \neq 0$ . From (4.8), we can see that the bias of  $\tilde{R}^{\text{local}}(t)$  depends on  $\alpha_0$ ,  $\alpha_1$  and the ratio of  $\mu^{\text{imported}}(t)$  and  $\mu^{\text{local}}(t)$ . When  $\mu^{\text{imported}}(t)/\mu^{\text{local}}(t) = 1$ , we have  $\tilde{R}^{\text{local}}(t) > R^{\text{local}}(t)$  if  $\alpha_0 > \alpha_1$ , and  $\tilde{R}^{\text{local}}(t) < R^{\text{local}}(t)$  if  $\alpha_0 < \alpha_1$ .

### 4.2.3 Bayesian hierarchical modeling to account for misidentification

We propose a Bayesian framework to estimate  $R^{\text{local}}(t)$  using noisy surveillance data. Following Fraser (2007); Thompson et al. (2019); Cori et al. (2013), we specify

$$I^{\text{local}}(t) | R^{\text{local}}(t), n(t-1), w(s) \sim \text{Pois}(R^{\text{local}}(t) \cdot \Lambda(t)), \text{ for } t > 0, \quad (4.9)$$

where  $\Lambda(t) = \sum_{s=1}^t w(s)I(t-s)$  is the total infectiousness of infected individuals at time  $t$ , and  $n(t-1)$  represent the historical data up to time  $t-1$  (i.e.,  $I^{\text{local}}(0), I^{\text{imported}}(0), \dots, I^{\text{local}}(t-1), I^{\text{imported}}(t-1)$ ). Note that  $\Lambda(t)$  is undefined for  $t=0$ . So, we assume that

$$I^{\text{local}}(0) | \mu^{\text{local}}(0) \sim \text{Pois}(\mu^{\text{local}}(0)). \quad (4.10)$$

And we assume the imported case counts follow a Poisson distribution:

$$I^{\text{imported}}(t) | \mu^{\text{imported}}(t) \sim \text{Pois}(\mu^{\text{imported}}(t)). \quad (4.11)$$

Next, we define relevant prior distributions. We assume a distribution for  $R^{\text{local}}(t)$  of the form

$$R^{\text{local}}(t) | n(t-1), w(s) \sim \text{Gamma}(a_{t|t-1}^{\text{local}}, b_{t|t-1}^{\text{local}}), \text{ for } t > 0. \quad (4.12)$$

This choice is similar to that in Thompson et al. (2019), but differs in that we specify gamma conditioned on the history, rather than marginally. The conditioning reflects the expectation that the evolution of  $R^{\text{local}}(t)$  is likely to depend on the course of infection in the population and intervention measures that may result. Analogously, we also assume gamma distributed priors for  $\mu^{\text{imported}}(t)$  and  $\mu^{\text{local}}(0)$ , that is,

$$\begin{aligned} \mu^{\text{imported}}(t) &\sim \text{Gamma}(a_t^{\text{imported}}, b_t^{\text{imported}}), \\ \mu^{\text{local}}(0) &\sim \text{Gamma}(a_0^{\text{local}}, b_0^{\text{local}}). \end{aligned} \quad (4.13)$$

In addition, we assume the convention that the misidentification rates are beta distributed, and hence given by

$$\begin{aligned} \alpha_0 &\sim \text{Beta}(\zeta_{\alpha_0}, \xi_{\alpha_0}), \\ \alpha_1 &\sim \text{Beta}(\zeta_{\alpha_1}, \xi_{\alpha_1}). \end{aligned} \quad (4.14)$$

By using Markov chain Monte Carlo (MCMC) simulation, we can get both estimates of  $R^{\text{local}}(t)$  and its uncertainty. We implement MCMC using the R package, NIMBLE (de Valpine et al. (2017, 2020a,b)) with the default assignment of sampler algorithms. The samplers assigned to the variables are as follows: Gibbs samplers are assigned to  $\mu^{\text{local}}(0)$  and  $\mu^{\text{imported}}(t)$ ,  $t \geq 0$ , which have conjugate relationships between their prior distribution and the distributions of their stochastic dependents; slice samplers

(Neal (2003)) are used for  $I^{\text{local}}(t)$  and  $I^{\text{imported}}(t)$ ,  $t \geq 0$ ; Metropolis-Hastings adaptive random-walk samplers are set to  $\alpha_0$ ,  $\alpha_1$  and  $R^{\text{local}}(t)$ ,  $t > 0$ .

#### 4.2.4 Estimating misidentification rates

Without any information on the misidentification rates, it is difficult to get an accurate estimator of  $R^{\text{local}}(t)$ . However, contact tracing data could provide adequate information to estimate the misidentification rates.

Let  $p_i$  be the probability that we think individual  $i$  is a local case based on the survey. Then,  $p_i$  can be modeled as a mixture of  $\alpha_0$  and  $1 - \alpha_1$ . Note that  $\alpha_1 \sim \text{Beta}(\zeta_{\alpha_1}, \xi_{\alpha_1})$  implies  $1 - \alpha_1 \sim \text{Beta}(\xi_{\alpha_1}, \zeta_{\alpha_1})$ . We thus model the distribution of  $p_i$  as a mixture of two beta distributions:

$$p_i \sim \pi_0 \text{Beta}(\zeta_{\alpha_0}, \xi_{\alpha_0}) + (1 - \pi_0) \text{Beta}(\xi_{\alpha_1}, \zeta_{\alpha_1}), \quad (4.15)$$

where  $\pi_0$  can be interpreted as the fraction of the diagnosed cases that are imported. By using the expectation-maximization (EM) algorithm, we can obtain estimators  $\hat{\zeta}_{\alpha_0}$ ,  $\hat{\xi}_{\alpha_0}$ ,  $\hat{\zeta}_{\alpha_1}$  and  $\hat{\xi}_{\alpha_1}$ .

Note that, if  $1 - \zeta_{\alpha_0}/(\zeta_{\alpha_0} + \xi_{\alpha_0}) - \zeta_{\alpha_1}/(\zeta_{\alpha_1} + \xi_{\alpha_1}) \neq 0$ , we obtain unbiased estimators of  $I^{\text{local}}(t)$  and  $I^{\text{imported}}(t)$

$$\begin{aligned} \hat{I}^{\text{local}}(t) &= \frac{[1 - \zeta_{\alpha_0}/(\zeta_{\alpha_0} + \xi_{\alpha_0})] \cdot \tilde{I}^{\text{local}}(t) - \zeta_{\alpha_0}/(\zeta_{\alpha_0} + \xi_{\alpha_0}) \cdot \tilde{I}^{\text{imported}}(t)}{1 - \zeta_{\alpha_0}/(\zeta_{\alpha_0} + \xi_{\alpha_0}) - \zeta_{\alpha_1}/(\zeta_{\alpha_1} + \xi_{\alpha_1})}, \\ \hat{I}^{\text{imported}}(t) &= \frac{[1 - \zeta_{\alpha_1}/(\zeta_{\alpha_1} + \xi_{\alpha_1})] \tilde{I}^{\text{imported}}(t) - \zeta_{\alpha_1}/(\zeta_{\alpha_1} + \xi_{\alpha_1}) \tilde{I}^{\text{local}}(t)}{1 - \zeta_{\alpha_0}/(\zeta_{\alpha_0} + \xi_{\alpha_0}) - \zeta_{\alpha_1}/(\zeta_{\alpha_1} + \xi_{\alpha_1})}. \end{aligned} \quad (4.16)$$

Thus, good initial values of  $I^{\text{local}}(t)$  and  $I^{\text{imported}}(t)$  in MCMC are estimators of  $\hat{I}^{\text{local}}(t)$  and  $\hat{I}^{\text{imported}}(t)$  based on the estimated misidentification rates, i.e., replacing  $\zeta_{\alpha_0}$ ,  $\xi_{\alpha_0}$ ,  $\zeta_{\alpha_1}$ ,  $\xi_{\alpha_1}$  in (4.16) by  $\hat{\zeta}_{\alpha_0}$ ,  $\hat{\xi}_{\alpha_0}$ ,  $\hat{\zeta}_{\alpha_1}$ ,  $\hat{\xi}_{\alpha_1}$ .

### 4.3 Results

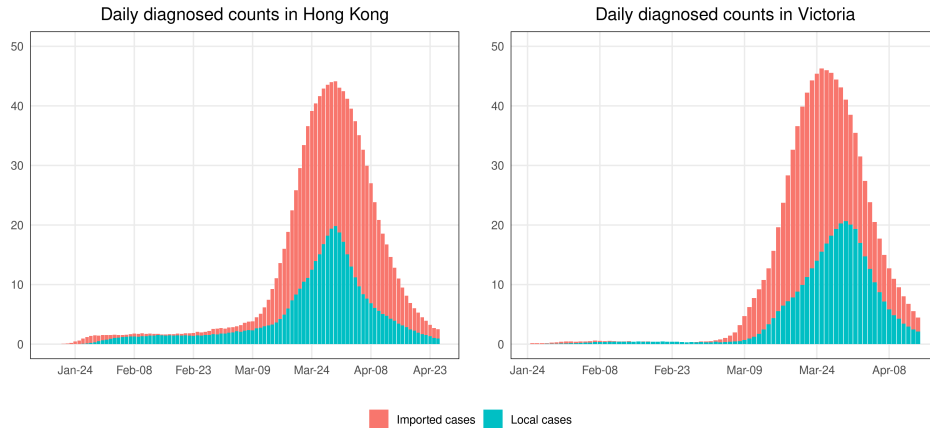
In this section, we conduct some simulations to illustrate the performance of the proposed estimation methods. And we apply our method to two real data sets. One is surveillance data of COVID-19 cases in Hong Kong that includes contact tracing information, including travel history data (Adam et al. (2020)). They collected information on 1,038 SARS-CoV-2 cases confirmed between 23 January and 28 April 2020. And they identified 355 local cases and 683 imported cases. The other data set is from the COVID-19 pandemic in Victoria, Australia, studied in Seemann et al. (2020). There they had 1,333 laboratory-confirmed cases of COVID-19 between 6 January and 14 April 2020. After excluding duplicate patients from cases, they identified 345 local cases and 558 imported cases.

We consider two settings, a simulation setting and an application setting. In the simulation setting, we first use surveillance data from Hong Kong and Victoria to create realistic simulated data, and then we add identification errors to the ‘true’ local and imported cases derived from the simulated epidemics, finally we estimate the local time-varying reproduction number using the noisy local and imported cases counts. In the application setting, we assume that identified local and imported cases in the real data sets are with some error. The former results allow us to understand what properties can be expected of our estimators, while the latter are reflective of what would be observed in practice with such data.

#### 4.3.1 Simulation study

In this simulation study, we used Covasim (Kerr et al. (2020)), a stochastic individual-based model for transmission of SARS-CoV-2, calibrated to the epidemics in Hong Kong and Victoria. Figure 4.1 shows the average daily local and imported diagnosed counts over 1,000 trials. The noisy  $\tilde{I}^{\text{local}}(t)$  and  $\tilde{I}^{\text{imported}}(t)$  are generated

according to (4.6). We set  $\alpha_0 \sim \text{Beta}(2, 18)$  (mean of 0.1), and  $\alpha_1 \sim \text{Beta}(2, 8)$  (mean of 0.2),  $\text{Beta}(4, 8)$  (mean of 0.33), or  $\text{Beta}(8, 8)$  (mean of 0.5) to see the effect of small  $\alpha_0$  and large  $\alpha_1$ . This might happen if the definition of imported cases relies on travel history collected in the case investigation and some people are infected locally, even though they have a travel history within 14 days prior to symptom onset. We also consider  $\alpha_1 \sim \text{Beta}(2, 18)$ , and  $\alpha_0 \sim \text{Beta}(2, 8)$ ,  $\text{Beta}(4, 8)$ , or  $\text{Beta}(8, 8)$  (corresponding to small  $\alpha_1$  and large  $\alpha_0$ , which might occur if cases are defined as local when we are not sure about their source of infection.) We assume that both  $\alpha_0$  and  $\alpha_1$  are unknown.



**Figure 4.1:** The means of daily local and imported diagnosed counts in 1,000 simulation trials for epidemics in Hong Kong and Victoria.

We evaluate the estimate for  $R^{\text{local}}(t)$  in terms of a corresponding posterior, and 95% credible intervals. Figure 4.2 and 4.3 show the simulation results, in which we run MCMC chains of 10,000 samples for each of 1,000 simulated epidemic trials. Figure 4.2 assumes that we are more likely to misclassify local cases as imported cases and Figure 4.3 assumes that we are more likely to misclassify imported cases as local cases. For comparison purposes, we compute  $R_*^{\text{local}}(t)$  and  $R^{\text{local}}(t)$  defined in (4.1) and (4.2) by approximating  $\mu_*^{\text{local}}(t)$ ,  $\mu_*(t)$ ,  $g(s)$ ,  $\mu^{\text{local}}(t)$ ,  $\mu(t)$ ,  $w(s)$  using 1,000 simulation trials. And we calculate the widely used estimator of  $\tilde{R}^{\text{local}}(t)$  defined in (4.3), which

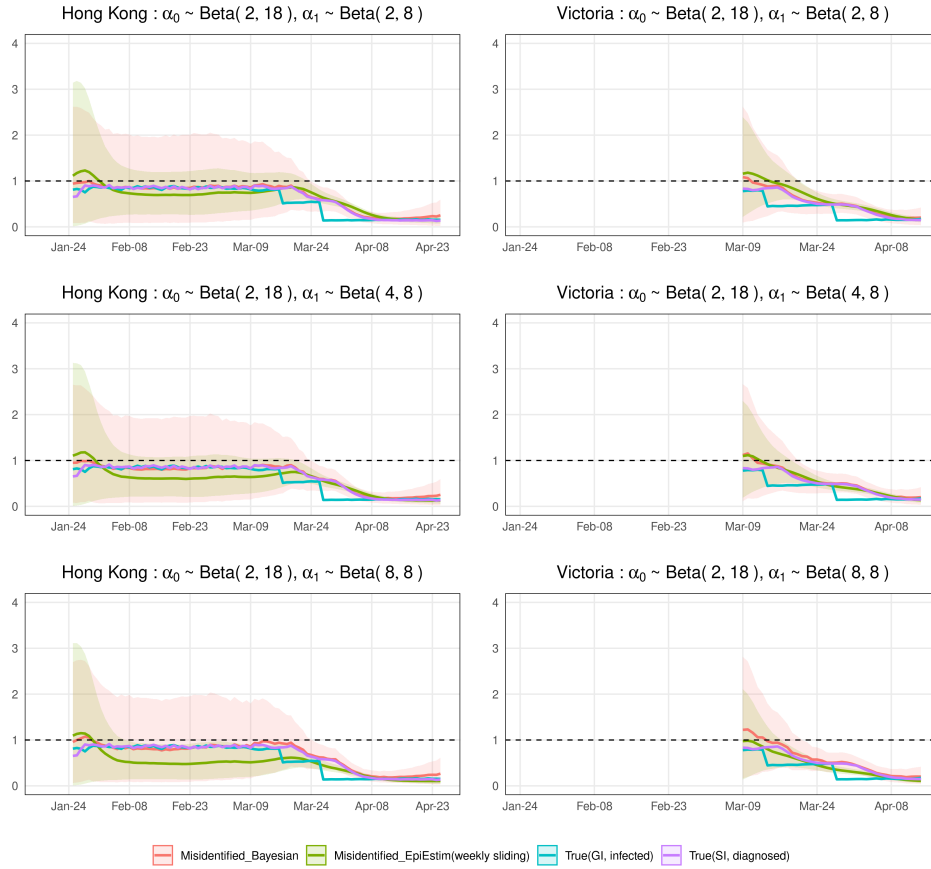
is implemented in the R package, EpiEstim (Cori et al. (2020)). We view it as a representative estimator that does not account for misidentification, i.e., it treats the noisy local and imported cases as true.

In the simulated epidemics for both Hong Kong and Victoria, if we ignore the misidentification, we will underestimate  $R^{\text{local}}(t)$  when the mean of  $\alpha_0$  is small and the mean of  $\alpha_1$  is relatively large (Figure 4.2), and overestimate  $R^{\text{local}}(t)$  when the mean of  $\alpha_1$  is small and the mean of  $\alpha_0$  is relatively large (Figure 4.3), with the biases increasing when the means of  $\alpha_0$  and  $\alpha_1$  increase. The results are consistent with (4.8) implying that the biases will lead to inappropriate public health response, i.e., inadequate interventions or overreaction. We correct the bias by our Bayesian hierarchical framework. The biases of our estimators are close to zero in all cases. The 95% credible intervals of our estimators are wide in the first two months because the number of incident cases are very low. For the last month or so when the diagnosed counts are relatively high, the 95% credible intervals are narrow.

### 4.3.2 Application

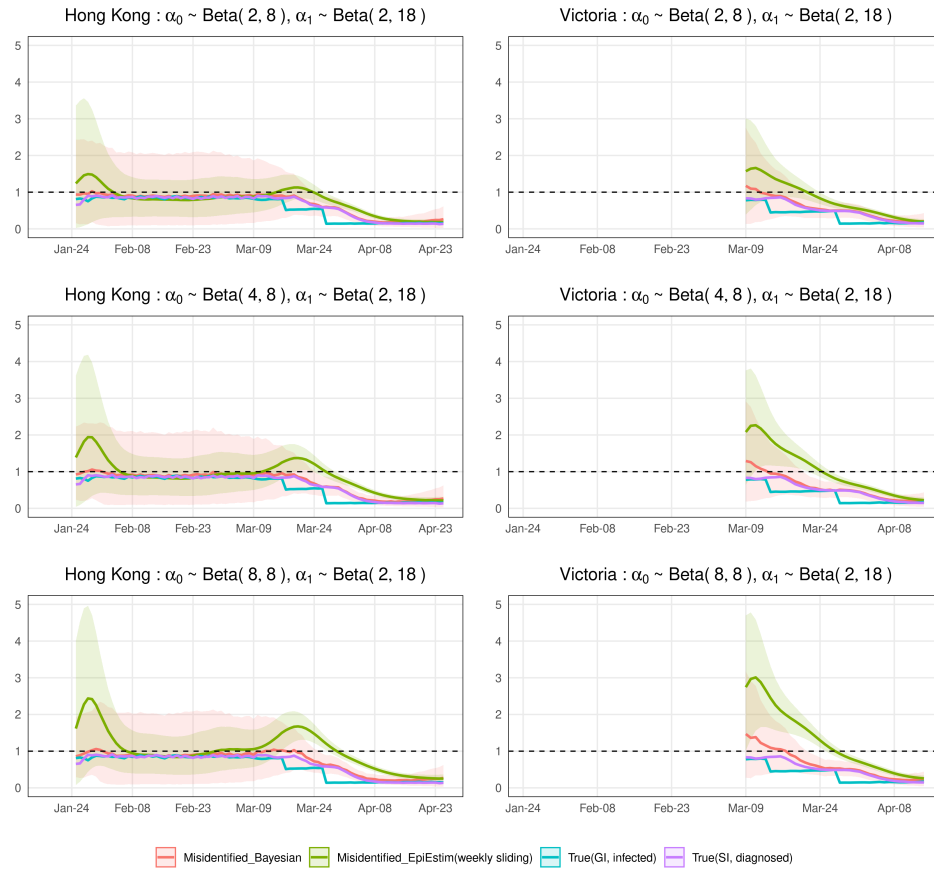
We apply our proposed methods to surveillance data of COVID-19 cases in Hong Kong and Victoria. Figure 4.4 (a) and (b) show the daily local and imported cases counts in Hong Kong and Victoria. For Hong Kong data, Adam et al. (2020) calculated the serial intervals using a gamma distribution and estimated shape and rate parameters of 2.23 and 0.37, respectively (corresponding to a mean of around 6 days and standard deviation of around 4 days). There is no specific serial interval that has been calculated for Victoria. Considering the epidemic curve in Victoria is relatively similar to that in Hong Kong, we use the same serial interval distribution when we estimate  $R^{\text{local}}(t)$  in Victoria.

Figure 4.4 (c) and (d) show estimates for  $R^{\text{local}}(t)$  under three assumed scenarios: 1) no identification error, 2) small  $\alpha_0$  and large  $\alpha_1$ , 3) small  $\alpha_1$  and large  $\alpha_0$ . We run



**Figure 4.2:** Estimations of local time-varying reproduction numbers in simulated epidemics for Hong Kong and Victoria under three sets of error misidentification rates:  $\alpha_0 \sim \text{Beta}(2, 18)$ , and  $\alpha_1 \sim \text{Beta}(2, 8)$ ,  $\text{Beta}(4, 8)$ , or  $\text{Beta}(8, 8)$ . The error bands are the averages of 95% credible intervals over 1,000 trials. Note that the differences between the blue curve ( $R_*^{\text{local}}(t)$ ) and the purple curve ( $R^{\text{local}}(t)$ ) are due to the differences among infected dates, symptom onset dates, diagnosed dates.

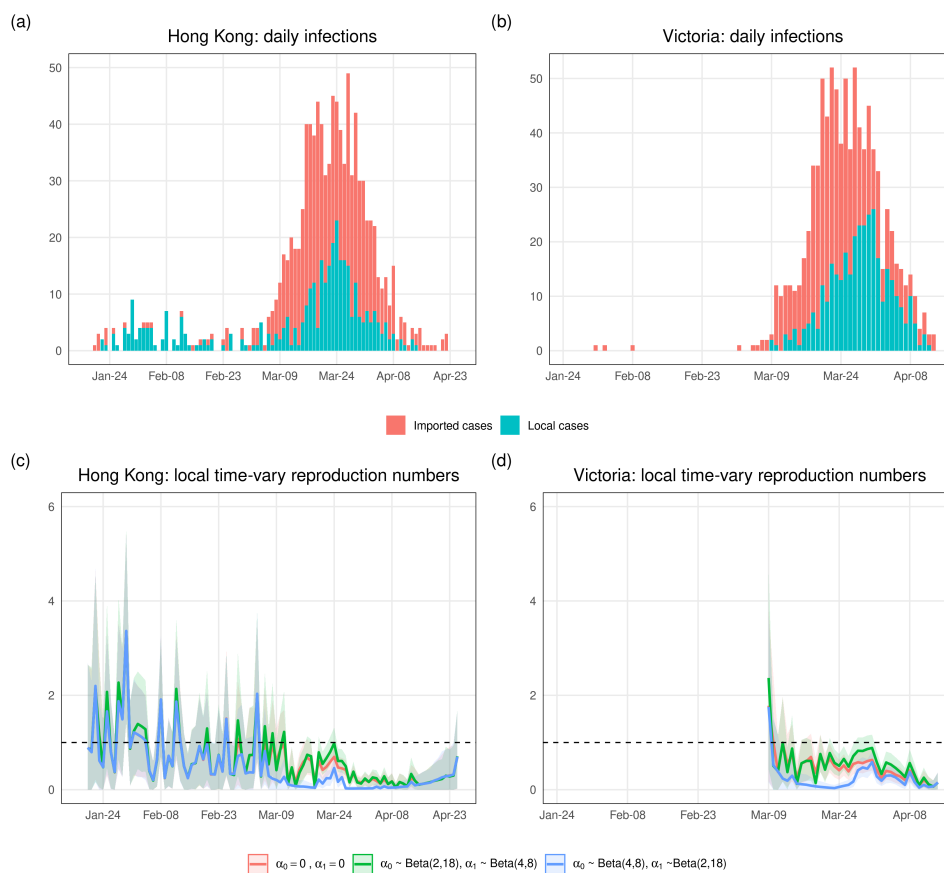
MCMC chains of 10,000 samples and the error bands are the 95% credible intervals. We can see that the estimated local time-varying reproduction numbers are quite different when the two identification error rates are about 10% and 30%. If we think we are more likely to misclassify local cases as imported, then we should trust the curve corresponding to scenario 2). If imported cases are more likely to be misidentified as local, then the curve corresponding to scenario 3) is reliable. And if we believe the



**Figure 4.3:** Estimations of local time-varying reproduction numbers in simulated epidemics for Hong Kong and Victoria under three sets of error misidentification rates:  $\alpha_1 \sim \text{Beta}(2, 18)$ , and  $\alpha_0 \sim \text{Beta}(2, 8)$ ,  $\text{Beta}(4, 8)$ , or  $\text{Beta}(8, 8)$ . The error bands are the averages of 95% credible intervals over 1,000 trials.

identification error is close to zero, we should trust the estimate under scenario 1).

Ultimately, we see that the ability to account for identification error appropriately in reporting the local time-varying reproduction number can lead to substantially different conclusions than use of the original, noisy local time-varying reproduction number. These differences can then in turn be translated to decision making for public health response.



**Figure 4.4:** Epidemic curves of COVID-19 cases and estimations of local time-varying reproduction numbers in Hong Kong and Victoria. (a) The epidemic curve of daily cases of laboratory-confirmed SARS-CoV-2 infection in Hong Kong by symptom onset date and colored by case category. Asymptomatic cases are included here by date of confirmation. (b) The epidemic curve of the coronavirus disease cases in Victoria by sample collection date and colored by case category. (c) and (d) Estimations of local time-varying reproduction numbers under three assumed scenarios: 1) no identification error, 2)  $\alpha_0 \sim \text{Beta}(2, 18)$  and  $\alpha_1 \sim \text{Beta}(4, 8)$  (around 10% imported cases are misclassified as local and around 33.3% local cases are misclassified as imported), 3)  $\alpha_0 \sim \text{Beta}(4, 8)$  and  $\alpha_1 \sim \text{Beta}(2, 18)$  (around 33.3% imported cases are misclassified as local and around 10% local cases are misclassified as imported). The bands are the 95% credible intervals.

## Data accessibility

No primary data are used in this chapter. Secondary data sources are taken from Adam et al. (2020); Seemann et al. (2020). These data and the code neces-

sary to reproduce the results in this chapter are available at <https://github.com/KolaczykResearch/EstimLocalRt>.

## Chapter 5

# Discussion

In this Chapter, we summarize the work in the dissertation and discuss future directions.

### 5.1 Inference for metrics for network-based epidemic surveillance

We have quantified the bias and variance of the observed branching factor in noisy networks and developed a general framework for estimation of the true branching factor in contexts wherein one has observations of noisy networks. Our approach requires as few as three replicates of network observations, and employs method-of-moments techniques to derive estimators and establish their asymptotic consistency and normality. Simulations demonstrate that substantial inferential accuracy by method-of-moments estimators is possible in networks of even modest size when nontrivial noise is present. And our application to contact networks in British secondary schools and a French hospital shows that the gains offered by our approach over presenting the observed branching factor can be pronounced.

We have pursued a frequentist approach to the problem of uncertainty quantification for the branching factor. If the replicates necessary for our approach are unavailable in a given setting, a Bayesian approach is a natural alternative. For example, posterior-predictive checks for goodness-of-fit based on examination of a handful of network summary measures is common practice (e.g., Bloem-Reddy and Orbanz (2018)). Note,

however, that the Bayesian approach requires careful modeling of the generative process underlying  $G$  and typically does not distinguish between signal and noise components. Our analysis is conditional on  $G$ , and hence does not require that  $G$  be modeled. It is effectively a ‘signal plus noise’ model, with the signal taken to be fixed but unknown. Related work has been done in the context of graphon modeling, with the goal of estimating network motif frequencies (e.g., Latouche and Robin (2016)). However, again, one typically does not distinguish between signal and noise components in this setting. Additionally, we note that the problem of practical graphon estimation itself is still a developing area of research.

Our work here sets the stage for extensions to various thresholds and statistics which depend on the branching factor. Recall that these include the epidemic threshold  $1/(\kappa - 1)$  and the immunization threshold  $1 - 1/(\lambda\kappa)$ , where  $\lambda$  is the spreading rate (Pastor-Satorras et al. (2015)). Replacing  $\kappa$  with  $\hat{\kappa}$ , we obtain asymptotically unbiased estimators for the corresponding thresholds. The asymptotic distributions can be derived from the delta method. In addition, the total branching factor of the network is important for epidemic spreading and immunization strategy in multiplex networks (e.g., Buono et al. (2014)).

Our choice to work with independent network noise is both natural and motivated by convenience. And our results of method-of-moments estimators still hold when there is some dependency across (non)edges. A precise characterization of the dependency is typically problem-specific and hence a topic for further investigation.

One future direction for this work is uncertainty quantification for the eigenvalues of the adjacency matrix. Because, in general, the epidemic threshold of a network is the inverse of the largest eigenvalue of the adjacency matrix (Pastor-Satorras et al. (2015)). The techniques for the eigenvalue of the adjacency matrix are different from those for the ratio of the second moment to the first moment of degree distribution.

Another more general direction is to pursue similar lines of research for other standard quantities associated with surveillance, including elements of the epidemic curve used in detecting emergence and peaks, and optimal surveillance groups (Herrera et al. (2016); Colman et al. (2019)).

## 5.2 Experiments on noisy networks

We have quantified biases and variances of standard estimators in noisy networks and developed a general framework for estimation of true average causal effects in contexts wherein one has observations of noisy networks. Our approach requires knowledge or consistent estimates of the corresponding noise parameters, the latter which can be obtained with as few as three replicates of network observations. We employ method-of-moments techniques to derive estimators and establish their asymptotic unbiasedness and consistency. Simulations in British secondary schools contact networks demonstrate that substantial inferential accuracy by method-of-moments estimators is possible in networks of even modest size when nontrivial noise is present.

Our work here sets the stage for extensions to other potential outcome frameworks and exposure models. Here we sketch the key elements of one such extension. For example, consider the exposure mapping  $f$  as following:

$$f(\mathbf{z}, \mathbf{A}_i) = \begin{cases} c_{11'}(\text{Direct} + \geq m_i \text{ Neighborhood Exposure}), & z_i I_{\{\mathbf{z}^\top \mathbf{A}_i > 0\}} \geq m_i, \\ c_{10'}(\text{Direct} + < m_i \text{ Neighborhood Exposure}), & z_i I_{\{\mathbf{z}^\top \mathbf{A}_i > 0\}} < m_i, \\ c_{01'}(\geq m_i \text{ Neighborhood Exposure}), & (1 - z_i) I_{\{\mathbf{z}^\top \mathbf{A}_i > 0\}} \geq m_i, \\ c_{00'}(< m_i \text{ Neighborhood Exposure}), & (1 - z_i) I_{\{\mathbf{z}^\top \mathbf{A}_i > 0\}} < m_i, \end{cases} \quad (5.1)$$

where  $m_i \geq 1$ . When  $m_i = 1$ , it reduces to (3.2). And if  $m_i = k$ , then the level  $c_{11'}$  is known as the absolute  $k$ -neighborhood exposure (Ugander et al. (2013)). When

$m_i = qd_i$ ,  $0 \leq q \leq 1$ , the level  $c_{11'}$  is called the fractional  $q$ -neighborhood exposure (Ugander et al. (2013)). The generalized four-level exposure model provides useful abstractions for the analysis of networked experiments. For example, infectious diseases (e.g., like COVID-19) are more likely to spread between people closely connected in a social network. And being in contact with more people with the disease means that, in theory, they are more likely to contract the disease.

As an illustration, suppose that treatment is assigned to the  $N_v$  individuals in a network through Bernoulli random sampling, with probability  $p$ . The exposure probabilities for four levels can be found in Appendix 3.5.3. Next, we show orders of the exposure probabilities for nodes with varying degrees in Theorem 5.1. If  $d_i = \Theta(1/p)$ , upper bounds for four exposure probabilities do not depend on  $m_i$ . When  $d_i = \omega(1/p)$ , upper bounds for  $p_i^e(c_{10'})$  and  $p_i^e(c_{00'})$  increase as  $m_i$  increases. And upper bounds for  $p_i^e(c_{11'})$  and  $p_i^e(c_{01'})$  decrease as  $m_i$  increases if  $d_i = o(1/p)$ . See supplementary material B.4 for the proof.

**Theorem 5.1** *Assume a generalized four-level exposure model and Bernoulli random assignment of treatment with  $p = o(1)$ . And for all  $i$ ,  $d_i \geq m_i$ . Then, for all integers  $m_i \geq 1$  and  $m_i = \mathcal{O}(1)$ , the orders of exposure probabilities are as in Table 5.1.*

	$p_i^e(c_{11'})$	$p_i^e(c_{10'})$	$p_i^e(c_{01'})$	$p_i^e(c_{00'})$
$d_i = \omega(1/p)$	$\mathcal{O}(p)$	$\mathcal{O}(p(d_i p)^{m_i-1}/e^{d_i p})$	$\mathcal{O}(1)$	$\mathcal{O}((d_i p)^{m_i-1}/e^{d_i p})$
$d_i = \Theta(1/p)$	$\mathcal{O}(p)$	$\mathcal{O}(p)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$
$d_i = o(1/p)$	$\mathcal{O}(p(d_i p)^{m_i})$	$\mathcal{O}(p)$	$\mathcal{O}((d_i p)^{m_i})$	$\mathcal{O}(1)$

Then, we can construct regularity conditions for the average causal effect estimators to be consistent. Similarly, one can quantify biases and variances of standard estimators in noisy networks and develop a general framework for estimation of true average causal effects. These require additional work due to the complexities of formulas for exposure probabilities.

Another interesting direction for this work is casual inference under directed networks or weighted networks. One idea for modeling noisy undirected weighted network is as following. Let  $w_{i,j}$  be the weight between node  $i$  and node  $j$  in the true network. For example, in a contact network,  $w_{i,j}$  can be defined as the frequency of contacts between node  $i$  and node  $j$ . Then, the marginal distribution of the observed weight in the noisy network has the form

$$\tilde{w}_{i,j} \sim \text{Bin}(w_{i,j}, 1 - \gamma_{i,j}) + \text{Pois}(\lambda_{i,j}),$$

where  $\gamma_{i,j}$  is the probability of missing one true contact,  $\lambda_{i,j}$  is the expected number of false contacts.

Our choice to work with independent network noise is both natural and motivated by convenience. A precise characterization of the dependency is typically problem-specific and hence a topic for further investigation.

### 5.3 Estimation of reproduction numbers from noisy surveillance data

We have developed a general framework for estimation of the true local time-varying reproduction numbers in contexts wherein one has identified local and imported case counts with some error. Simulations demonstrate that substantial inferential accuracy by our estimators is possible when nontrivial error is present. And our application to epidemics in Hong Kong and Victoria shows that the gains offered by our approach over presenting the noisy local instantaneous reproduction number can be pronounced.

We have shown examples on a country level, but our method could be useful for states, cities, or more local settings, such as a university trying to determine if there is substantial local transmission occurring. Our approach requires daily numbers of local and imported cases, serial interval, and contact tracing data or other data to

provide adequate information to estimate the misidentification rates.

We have pursued a Bayesian approach to the problem of estimating the local instantaneous reproduction number. The credible intervals are relatively wide when the number of cases is low. To improve the performance at low case incidence, Kalman filtering is a natural approach. Estimating the time-vary reproduction number by Kalman filtering is an emerging topic. For instance, Parag (2020) constructed a recursive Bayesian smoother for estimating the effective reproduction number from the incidence of an infectious disease in real time and retrospectively. However, one typically does not distinguish between local and imported cases in this setting.

The identification errors are informed by contact tracing survey data in our approach. If the data from the survey is categorical (e.g., we ask people where they were infected and attach some qualitative measure of our confidences that we think they are local cases), we can transform them into numerical values. For example, Patki et al. (2016) proposed a method that converts categorical variables to numerical data for Gaussian distribution. We could modify the method to convert categorical variables to Beta distributed data. If the survey data is unavailable, using genomic data is a natural alternative. Genomic surveillance has been used to detect transmission clusters and to provide information on the possible source of individual cases (Leavitt et al. (2020); Meredith et al. (2020); Deng et al. (2020); Poon et al. (2016); Sansone et al. (2020); Peters et al. (2016)).

We have showed the results of retrospective estimation. And it is computationally feasible to run MCMC on each day to obtain real time estimators; it takes about 5 minutes for the MCMC chain of 10,000 samples. To reduce the computational cost, one approach is adaptive MCMC methods (Haario et al. (2001); Roberts and Rosenthal (2007)), which use the covariance structure of the posterior distribution to design proposal distributions. Other methods include stochastic Newton (Martin

et al. (2012)) and Riemannian manifold MCMC (Girolami and Calderhead (2011)), which construct efficient proposals by local derivative information.

Due to the presence of heterogeneity in the level of connectivity of contact neighborhoods for most real-world contact networks, it is interesting to analyze disease transmission dynamics among different compartments (e.g., students, staff and faculty in an university). Consider a population with  $K$  compartments. Let  $R_{k,l}(t)$  be a time-varying reproduction number that parameterizes explicitly the rate at which individuals from compartment  $l$  infect those in compartment  $k$ . Let  $I_{k,l}(t)$  denote the number of true newly diagnose cases at time  $t$  for compartment  $k$  infected by people in compartment  $l$ . Then we can estimate  $R_{k,l}(t)$  using the surveillance data (including  $I_{k,l}(t)$ ) by Kalman filtering or a similar approach in Thompson et al. (2019). To account for noise in the surveillance data, a simple model is

$$\begin{aligned} & \tilde{I}_{k,l}(t) | I_{k,1}(t), \dots, I_{k,K}(t), I_k^{\text{imported}}(t), \beta_0, \beta_1 \\ & \sim \text{Bin}(I_{k,l}(t), 1 - \beta_1) + \text{Bin}(I_k^{\text{imported}}(t) + \sum_{j \in \{1, \dots, K\} \setminus l} I_{k,j}(t), \beta_0), \end{aligned} \quad (5.2)$$

where  $I_k^{\text{imported}}(t)$  is the number of newly diagnose imported cases at time  $t$  for compartment  $k$ ,  $\beta_0$  denotes the probability that an imported case or a case having index case in compartment  $j$  ( $j \neq l$ ) is misidentified as a case infected by people in compartment  $l$ , and  $\beta_1$  is the probability that a case infected by people in compartment  $l$  is misidentified as a case having other index cases. In addition, if  $\tilde{I}_{k,l}(t)$  is unavailable, one method to estimate  $R_{k,l}(t)$  is using the network connectivity information (i.e., within-compartments and between-compartments contacts).

## Appendix A

# Supplementary Materials to Chapter 2

We provide theorems and corollaries for asymptotic bias and variance of the observed branching number and the proofs, proofs of theorems for method-of-moments estimator  $\hat{\kappa}$ , and the algorithm for estimation of asymptotic variance of  $\hat{\kappa}$ .

### A.1 Theorems and corollaries for bias of the observed branching number

In this section, we first quantify the asymptotic bias of the observed branching factor for arbitrary true networks. We then show specific results for four typical classes of networks: sparse and homogeneous, sparse and inhomogeneous, dense and homogeneous, and dense and inhomogeneous. Note that Corollary 2.1 in Chapter 2 corresponds to Corollary A.1 and Corollary A.3, and Corollary 2.2 in Chapter 2 corresponds to Corollary A.2 and Corollary A.4.

**Theorem A.1** *We define  $X = \sum_{i=1}^{N_v} \tilde{d}_i^2$ ,  $Y = \sum_{i=1}^{N_v} \tilde{d}_i$  and we assume  $\mathbb{E}Y > 0$ , and  $\mathbb{E}Y = \Omega(N_v)$  ( $N_v \rightarrow \infty$ ). Then, under Assumption 2.2, for any  $\eta > 0$ , we have*

$$\text{Bias}[\tilde{\kappa}] = \frac{\mathbb{E}X}{\mathbb{E}Y} - \kappa + \mathcal{O}\left(\frac{1}{(\mathbb{E}Y)^{1/(2+\eta)}} \frac{\mathbb{E}X}{\mathbb{E}Y}\right) \text{ as } N_v \rightarrow \infty.$$

**Theorem A.2** *Under assumptions in Theorem A.1 and Assumption 2.1 and 2.4, for any  $\eta > 0$ , we have*

$$\text{Bias}[\tilde{\kappa}] = (2 - \alpha - \beta) \left[ \alpha(N_v - 1) + \beta - (\alpha + \beta)\kappa \right] + \mathcal{O}\left(\frac{1}{(\mathbb{E}Y)^{1/(2+\eta)}} \frac{\mathbb{E}X}{\mathbb{E}Y}\right)$$

as  $N_v \rightarrow \infty$ .

**Corollary A.1 (Sparse and homogeneous)** *In the sparse homogeneous graph, where the average degree  $\bar{d} = \Theta(\log N_v)$  and the asymptotic degree distribution is the Poisson distribution with mean  $\bar{d}$ , under the assumptions in Theorem A.2 and  $\beta = \mathcal{O}(1)$  ( $N_v \rightarrow \infty$ ), for any  $\eta > 0$ , we have*

$$\text{Bias}[\tilde{\kappa}] = \mathcal{O}\left(\frac{\log N_v}{(N_v \log N_v)^{1/(2+\eta)}}\right) \text{ as } N_v \rightarrow \infty,$$

where  $\kappa = \Theta(\log N_v)$ .

**Corollary A.2 (Sparse and inhomogeneous)** *In the sparse inhomogeneous graph where the average degree  $\bar{d} = \Theta(\log N_v)$  and the asymptotic degree distribution is truncated Pareto distribution with shape  $\zeta$ , lower bound  $d_L$  and upper bound  $N_v - 1$ , under the assumptions in Theorem A.2 and  $\beta = \mathcal{O}(1)$  ( $N_v \rightarrow \infty$ ), for any  $\eta > 0$ , we have*

$$\text{Bias}[\tilde{\kappa}] = \begin{cases} -\beta(2 - \alpha - \beta)\kappa + \mathcal{O}\left(\max\left\{\log N_v, \frac{\kappa}{(N_v \log N_v)^{1/(2+\eta)}}\right\}\right) & \text{if } 0 < \zeta \leq 2 \\ -\beta(2 - \alpha - \beta)\frac{\kappa}{(\zeta - 1)^2} + \mathcal{O}(1) & \text{if } \zeta > 2 \end{cases}$$

as  $N_v \rightarrow \infty$ , where

$$\kappa = \begin{cases} \Theta(N_v), & \text{if } 0 < \zeta < 1 \\ \Theta(N_v / \log N_v), & \text{if } \zeta = 1 \\ \Theta(N_v^{2-\zeta} \cdot \log^{\zeta-1} N_v), & \text{if } 1 < \zeta < 2 \\ \Theta(\log^2 N_v), & \text{if } \zeta = 2 \\ \Theta(\log N_v), & \text{if } \zeta > 2. \end{cases}$$

**Remark A.1** *In Corollary A.2, by the definition of expectation,  $\zeta$ ,  $d_L$ ,  $\bar{d}$  and  $N_v$  satisfy the equation*

$$\bar{d} = \int_{d_L}^{N_v-1} x \cdot \frac{\zeta d_L^\zeta}{1 - \left(\frac{d_L}{N_v-1}\right)^\zeta} x^{-(\zeta+1)} dx.$$

*Under the condition  $\bar{d} = \Theta(\log N_v)$ , the relationship among them can be simplified. Similar relationships also hold in Corollary A.4, A.6 and A.8.*

Note that the  $\mathcal{O}$  term in Corollary A.2 is dominated by the corresponding  $\kappa$  asymptotically, so  $\text{Bias}(\tilde{\kappa}) = \Theta(\kappa)$ , reflecting the challenges of estimating  $\kappa$  in under heterogeneous degree distributions. In contrast,  $\text{Bias}(\tilde{\kappa}) = o(\kappa)$  in Corollary A.1.

**Corollary A.3 (Dense and homogeneous)** *In the dense homogeneous graph where the average degree  $\bar{d} = \Theta(N_v^c)$ ,  $0 < c < 1$ , and the asymptotic degree distribution is the Poisson distribution with mean  $\bar{d}$ , under the assumptions in Theorem A.2 and  $\beta = \mathcal{O}(1)$  ( $N_v \rightarrow \infty$ ), for any  $\eta > 0$ , we have*

$$\text{Bias}[\tilde{\kappa}] = \mathcal{O}\left(N_v^{c-\frac{c+1}{2+\eta}}\right) \text{ as } N_v \rightarrow \infty,$$

where  $\kappa = \Theta(N_v^c)$ .

**Corollary A.4 (Dense and inhomogeneous)** *In the dense inhomogeneous graph where the average degree  $\bar{d} = \Theta(N_v^c)$ ,  $0 < c < 1$ , and the asymptotic degree distribution is truncated Pareto distribution with shape  $\zeta$ , lower bound  $d_L$  and upper bound  $N_v - 1$ , under the assumptions in Theorem A.2 and  $\beta = \mathcal{O}(1)$  ( $N_v \rightarrow \infty$ ), for any  $\eta > 0$ , we have*

$$\text{Bias}[\tilde{\kappa}] = \begin{cases} -\beta(2 - \alpha - \beta)\kappa + \mathcal{O}\left(\max\left\{N_v^c, \frac{\kappa}{N_v^{(c+1)/(2+\eta)}}\right\}\right) & \text{if } 0 < \zeta \leq 2 \\ -\beta(2 - \alpha - \beta)\frac{\kappa}{(\zeta - 1)^2} + \mathcal{O}(\max\{N_v^{2c-1}, 1\}) & \text{if } \zeta > 2 \end{cases}$$

as  $N_v \rightarrow \infty$ , where

$$\kappa = \begin{cases} \Theta(N_v), & \text{if } 0 < \zeta < 1 \\ \Theta(N_v / \log N_v), & \text{if } \zeta = 1 \\ \Theta(N_v^{2-\zeta+c(\zeta-1)}), & \text{if } 1 < \zeta < 2 \\ \Theta(N_v^c \cdot \log N_v), & \text{if } \zeta = 2 \\ \Theta(N_v^c), & \text{if } \zeta > 2. \end{cases}$$

Note that the  $\mathcal{O}$  term in Corollary A.4 is dominated by the corresponding  $\kappa$  asymptotically, so  $\text{Bias}(\tilde{\kappa}) = \Theta(\kappa)$ . In contrast,  $\text{Bias}(\tilde{\kappa}) = o(\kappa)$  in Corollary A.3.

## A.2 Theorems and corollaries for variance of the observed branching number

In this section, we first compute the asymptotic variance of the observed branching factor for arbitrary true networks. We then show specific results for the same four types of networks as in Section A.1. Note that Theorem 2.3 in Chapter 2 corresponds to Corollary A.5 - A.8.

**Theorem A.3** *We assume  $\mathbb{E}Y > 0$ , and  $\mathbb{E}Y = \Omega(N_v)$  ( $N_v \rightarrow \infty$ ). Then, under Assumption 2,*

(i)

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}\left(\max\left\{\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right], \mathbb{P}(Y = 0) \cdot \left[\frac{\mathbb{E}X}{\mathbb{E}Y}\right]^2\right\}\right)$$

as  $N_v \rightarrow \infty$ .

(ii) For any  $\eta, \lambda > 0$ ,

$$\text{Var}[\tilde{\kappa}] = \mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] + \mathcal{O}\left(\max\left\{(\mathbb{E}Y)^{-1/(2+\eta)} \cdot \mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right], (\mathbb{E}Y)^{-2/(2+\lambda)} \cdot \left[\frac{\mathbb{E}X}{\mathbb{E}Y}\right]^2, \mathbb{P}(Y = 0) \cdot \left[\frac{\mathbb{E}X}{\mathbb{E}Y}\right]^2\right\}\right)$$

as  $N_v \rightarrow \infty$ .

**Theorem A.4** *Under the assumptions in Theorem A.3, Assumption 1 and 4, and  $1 - \beta = \Omega(N_v)$  ( $N_v \rightarrow \infty$ ),*

(i)

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}\left(\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right]\right) \text{ as } N_v \rightarrow \infty.$$

(ii) For any  $\eta, \kappa > 0$ ,

$$\begin{aligned} \text{Var}[\tilde{\kappa}] &= \mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] \\ &+ \mathcal{O}\left(\max\left\{(\mathbb{E}Y)^{-1/(2+\eta)} \cdot \mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right], (\mathbb{E}Y)^{-2/(2+\lambda)} \cdot \left[\frac{\mathbb{E}X}{\mathbb{E}Y}\right]^2\right\}\right) \end{aligned}$$

as  $N_v \rightarrow \infty$ .

Theorem A.3 (i) and Theorem A.4 (i) provide upper bounds for variances of the observed branching factors. And Theorem A.3 (ii) and Theorem A.4 (ii) derive good approximations of variances if the  $\mathcal{O}$  terms are dominated by the corresponding first terms asymptotically.

**Corollary A.5 (Sparse and homogeneous)** *In the sparse homogeneous graph where the average degree  $\bar{d} = \Theta(\log N_v)$  and the asymptotic degree distribution is the Poisson distribution with mean  $\bar{d}$ , under the assumptions in Theorem A.2 and  $\beta = \mathcal{O}(1)$  ( $N_v \rightarrow \infty$ ), we have*

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}\left(\left(\frac{\log N_v}{N_v}\right)^{1/2}\right) \text{ as } N_v \rightarrow \infty.$$

**Corollary A.6 (Sparse and inhomogeneous)** *In the sparse inhomogeneous graph where the average degree  $\bar{d} = \Theta(\log N_v)$  and the asymptotic degree distribution is truncated Pareto distribution with shape  $\zeta$ , lower bound  $d_L$  and upper bound  $N_v - 1$ , under the assumptions in Theorem A.2 and  $\beta = \mathcal{O}(1)$  ( $N_v \rightarrow \infty$ ), we have*

$$\text{Var}[\tilde{\kappa}] = \begin{cases} \mathcal{O}(N_v/\log N_v), & 0 < \zeta < 1 \\ \mathcal{O}(N_v/\log^2 N_v), & \zeta = 1 \\ \mathcal{O}((N_v/\log N_v)^{2-\zeta}), & 1 < \zeta < 5/2 \\ \mathcal{O}((\log N_v/N_v)^{1/2}), & \zeta \geq 5/2 \end{cases}$$

as  $N_v \rightarrow \infty$ .

**Corollary A.7 (Dense and homogeneous)** *In the dense homogeneous graph where the average degree  $\bar{d} = \Theta(N_v^c)$ ,  $0 < c < 1$ , and the asymptotic degree distribution is the Poisson distribution with mean  $\bar{d}$ , under the assumptions in Theorem*

A.2 and  $\beta = \mathcal{O}(1)$  ( $N_v \rightarrow \infty$ ), we have

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}(N_v^{(c-1)/2}) \text{ as } N_v \rightarrow \infty.$$

**Corollary A.8 (Dense and inhomogeneous)** *In the dense inhomogeneous graph where the average degree  $\bar{d} = \Theta(N_v^c)$ ,  $0 < c < 1$ , and the asymptotic degree distribution is truncated Pareto distribution with shape  $\zeta$ , lower bound  $d_L$  and upper bound  $N_v - 1$ , under the assumptions in Theorem A.2 and  $\beta = \mathcal{O}(1)$  ( $N_v \rightarrow \infty$ ), we have*

$$\text{Var}[\tilde{\kappa}] = \begin{cases} \mathcal{O}(N_v^{1-c}), & 0 < \zeta < 1 \\ \mathcal{O}(N_v^{1-c}/\log N_v), & \zeta = 1 \\ \mathcal{O}(N_v^{(2-\zeta)(1-c)}), & 1 < \zeta < 5/2 \\ \mathcal{O}(N_v^{(c-1)/2}), & \zeta \geq 5/2 \end{cases}$$

as  $N_v \rightarrow \infty$ .

Note that the orders of the variances are asymptotically dominated by the corresponding biases for all four cases. Therefore, in noisy contact networks, bias would appear to be the primary concern for the observed branching factor. The  $\mathcal{O}$  notations for variances in the homogeneous networks are bounded above by those in the inhomogeneous networks of the same network density.

### A.3 Proofs of theorems for bias of the observed branching number

#### Proof of Theorem A.1

Recall  $X = \sum_i \tilde{d}_i^2$  and  $Y = \sum_i \tilde{d}_i$ . Note that

$$\text{Bias}[\tilde{\kappa}] = \frac{\mathbb{E}X}{\mathbb{E}Y} - \kappa + \mathcal{O}\left(\frac{1}{(\mathbb{E}Y)^{1/(2+\eta)}} \frac{\mathbb{E}X}{\mathbb{E}Y}\right)$$

is equivalent to

$$\mathbb{E}[\tilde{\kappa}] - \frac{\mathbb{E}X}{\mathbb{E}Y} = \mathcal{O}\left(\frac{1}{(\mathbb{E}Y)^{1/(2+\eta)}} \frac{\mathbb{E}X}{\mathbb{E}Y}\right). \quad (\text{A.1})$$

By Jensen's inequality, we have

$$\left| \mathbb{E}[\tilde{\kappa}] - \frac{\mathbb{E}X}{\mathbb{E}Y} \right| = \frac{1}{\mathbb{E}Y} \left| \mathbb{E}\left[\frac{X(\mathbb{E}Y - Y)}{Y} \cdot I_{\{Y>0\}}\right] \right| \leq \frac{1}{\mathbb{E}Y} \cdot \mathbb{E}\left[\frac{X|\mathbb{E}Y - Y|}{Y} \cdot I_{\{Y>0\}}\right]. \quad (\text{A.2})$$

Then by additivity of expectation, for  $0 < \delta < 1$ ,  $\mathbb{E}\left[\frac{X|\mathbb{E}Y - Y|}{Y} \cdot I_{\{Y>0\}}\right]$  in (A.2) equals

$$\mathbb{E}\left[\frac{X|\mathbb{E}Y - Y|}{Y} \cdot I_{\{Y>0\}} \cdot I_{\{|Y - \mathbb{E}Y| \geq \delta \mathbb{E}Y\}}\right] + \mathbb{E}\left[\frac{X|\mathbb{E}Y - Y|}{Y} \cdot I_{\{Y>0\}} \cdot I_{\{|Y - \mathbb{E}Y| < \delta \mathbb{E}Y\}}\right]. \quad (\text{A.3})$$

Next, we find the upper bounds of two terms in (A.3). For the first term, by definitions of  $X$  and  $Y$ ,  $X/Y \cdot I_{\{Y>0\}} < N_v$  and  $|\mathbb{E}Y - Y| < N_v^2$ . Thus, we have

$$\mathbb{E}\left[\frac{X|\mathbb{E}Y - Y|}{Y} \cdot I_{\{Y>0\}} \cdot I_{\{|Y - \mathbb{E}Y| \geq \delta \mathbb{E}Y\}}\right] < N_v^3 \cdot \Pr(|Y - \mathbb{E}Y| \geq \delta \mathbb{E}Y).$$

Then, by Chernoff Bound, we obtain

$$\mathbb{E}\left[\frac{X|\mathbb{E}Y - Y|}{Y} \cdot I_{\{Y>0\}} \cdot I_{\{|Y - \mathbb{E}Y| \geq \delta \mathbb{E}Y\}}\right] < 2N_v^3 \cdot \exp\left(-\frac{\delta^2 \cdot \mathbb{E}Y}{6}\right).$$

For the second term, when  $|Y - \mathbb{E}Y| < \delta \mathbb{E}Y$ ,  $Y > (1 - \delta)\mathbb{E}Y$ . So, we obtain

$$\mathbb{E}\left[\frac{X|\mathbb{E}Y - Y|}{Y} \cdot I_{\{Y>0\}} \cdot I_{\{|Y - \mathbb{E}Y| < \delta \mathbb{E}Y\}}\right] < \frac{\delta}{(1 - \delta)} \cdot \mathbb{E}X.$$

By (A.2), we show

$$\left| \mathbb{E}[\tilde{\kappa}] - \frac{\mathbb{E}X}{\mathbb{E}Y} \right| \leq \frac{2N_v^3}{\mathbb{E}Y} \cdot \exp\left(-\frac{\delta^2 \cdot \mathbb{E}Y}{6}\right) + \frac{\delta}{(1 - \delta)} \cdot \frac{\mathbb{E}X}{\mathbb{E}Y}. \quad (\text{A.4})$$

Let  $L_1(N_v)$  and  $L_2(N_v)$  denote two terms on the right sides in (A.4). We choose  $\delta = (\mathbb{E}Y)^{-1/(2+\eta)}$ ,  $\eta > 0$ , such that

$$L_1(N_v) = o\left(\frac{\mathbb{E}X}{\mathbb{E}Y}\right) \text{ and } L_2(N_v) = o\left(\frac{\mathbb{E}X}{\mathbb{E}Y}\right) \text{ as } N_v \rightarrow \infty,$$

under the assumption  $\mathbb{E}Y = \Omega(N_v)$  as  $N_v \rightarrow \infty$ ,  $1 - \delta = \mathcal{O}(1)$ . By L'Hopital's rule, we have

$$L_1(N_v) = o(L_2(N_v)) \text{ as } N_v \rightarrow \infty.$$

These imply (A.1).

### Proof of Theorem A.2

We compute  $\mathbb{E}Y$  and  $\mathbb{E}X$  under Assumption 2.1 and 2.2,

$$\begin{aligned} \mathbb{E}Y &= \sum_{i=1}^{N_v} \mathbb{E}[\tilde{d}_i] = \sum_{i=1}^{N_v} \alpha(N_v - 1 - d_i) + (1 - \beta)d_i \\ &= \alpha N_v(N_v - 1) + (1 - \alpha - \beta) \sum_{i=1}^{N_v} d_i, \text{ and} \\ \mathbb{E}X &= \sum_{i=1}^{N_v} \mathbb{E}[\tilde{d}_i^2] = \sum_{i=1}^{N_v} (\text{var}[\tilde{d}_i] + (\mathbb{E}[\tilde{d}_i])^2) \\ &= \sum_{i=1}^{N_v} \left( \alpha(1 - \alpha)(N_v - 1 - d_i) \right. \\ &\quad \left. + \beta(1 - \beta)d_i + [\alpha(N_v - 1 - d_i) + (1 - \beta)d_i]^2 \right) \\ &= (1 - \alpha - \beta)^2 \sum_{i=1}^{N_v} d_i^2 + [\beta(1 - \beta) - \alpha(1 - \alpha) \\ &\quad + 2\alpha(N_v - 1)(1 - \alpha - \beta)] \sum_{i=1}^{N_v} d_i + \alpha N_v(N_v - 1)[1 - \alpha + \alpha(N_v - 1)]. \end{aligned} \quad (\text{A.5})$$

Then, under Assumption 2.4, (A.5) leads to

$$\begin{aligned}\mathbb{E}Y &= \sum_{i=1}^{N_v} d_i, \\ \mathbb{E}X &= (1 - \alpha - \beta)^2 \sum_{i=1}^{N_v} d_i^2 + (2 - \alpha - \beta) [\alpha(N_v - 1) + \beta] \sum_{i=1}^{N_v} d_i.\end{aligned}$$

Plugging the value of  $\mathbb{E}X/\mathbb{E}Y$  into the bias expression in Theorem A.1 completes the proof.

## A.4 Proofs of corollaries for bias of the observed branching number

### Proof of Corollary A.1

By homogeneity, we obtain

$$\sum_{i=1}^{N_v} d_i^2 = (\bar{d} + 1)\bar{d}N_v.$$

By edge unbiasedness, we have

$$\alpha = \frac{\beta\bar{d}}{N_v - 1 - \bar{d}}. \tag{A.6}$$

Thus,

$$\alpha(N_v - 1) + \beta - (\alpha + \beta) \frac{\sum_{i=1}^{N_v} d_i^2}{\sum_{i=1}^{N_v} d_i} = -\alpha, \tag{A.7}$$

and

$$\frac{\mathbb{E}X}{\mathbb{E}Y} = \bar{d} + 1 - \alpha(2 - \alpha - \beta).$$

By Theorem A.2, for any  $\eta > 0$ , we have

$$\text{Bias}[\tilde{\kappa}] = \mathcal{O}\left(\frac{\log N_v}{(N_v \log N_v)^{1/(2+\eta)}}\right) \text{ as } N_v \rightarrow \infty.$$

### Proof of Corollary A.2

First, we compute the first and second moments of truncated Pareto distribution.

$$\int_{d_L}^{N_v-1} x \cdot \frac{\zeta d_L^\zeta}{1 - \left(\frac{d_L}{N_v-1}\right)^\zeta} x^{-(\zeta+1)} dx = \begin{cases} \zeta d_L^\zeta \cdot \frac{\log\left(\frac{N_v-1}{d_L}\right)}{1 - \left(\frac{d_L}{N_v-1}\right)^\zeta}, & \text{if } \zeta = 1 \\ \frac{\zeta d_L}{\zeta - 1} \cdot \frac{1 - \left(\frac{d_L}{N_v-1}\right)^{\zeta-1}}{1 - \left(\frac{d_L}{N_v-1}\right)^\zeta}, & \text{otherwise,} \end{cases}$$

$$\int_{d_L}^{N_v-1} x^2 \cdot \frac{\zeta d_L^\zeta}{1 - \left(\frac{d_L}{N_v-1}\right)^\zeta} x^{-(\zeta+1)} dx = \begin{cases} \zeta d_L^\zeta \cdot \frac{\log\left(\frac{N_v-1}{d_L}\right)}{1 - \left(\frac{d_L}{N_v-1}\right)^\zeta}, & \text{if } \zeta = 2 \\ \frac{\zeta d_L^2}{\zeta - 2} \cdot \frac{1 - \left(\frac{d_L}{N_v-1}\right)^{\zeta-2}}{1 - \left(\frac{d_L}{N_v-1}\right)^\zeta}, & \text{otherwise.} \end{cases}$$

(A.8)

Note that  $\bar{d} = \sum_{i=1}^{N_v} d_i / N_v = \Theta(\log N_v)$ . So, as  $N_v \rightarrow \infty$ , we obtain

$$d_L \sim \begin{cases} \left(\frac{1-\zeta}{\zeta} \cdot \frac{\bar{d}}{N_v^{1-\zeta}}\right)^{1/\zeta}, & \text{if } 0 < \zeta < 1 \\ \bar{d} / \log N_v, & \text{if } \zeta = 1 \\ \frac{\zeta-1}{\zeta} \bar{d}, & \text{if } \zeta > 1. \end{cases}$$

(A.9)

Thus, as  $N_v \rightarrow \infty$ , we have

$$\frac{\sum_{i=1}^{N_v} d_i^2}{\sum_{i=1}^{N_v} d_i} \sim \begin{cases} \frac{1-\zeta}{2-\zeta} N_v, & \text{if } 0 < \zeta < 1 \\ N_v / \log N_v, & \text{if } \zeta = 1 \\ \frac{\zeta-1}{2-\zeta} \left(\frac{\zeta-1}{\zeta}\right)^{\zeta-1} N_v^{2-\zeta} \cdot (\bar{d})^{\zeta-1}, & \text{if } 1 < \zeta < 2 \\ \bar{d} \cdot \log N_v / 2, & \text{if } \zeta = 2 \\ \frac{(\zeta-1)^2}{\zeta(\zeta-2)} \bar{d}, & \text{if } \zeta > 2. \end{cases} \quad (\text{A.10})$$

By edge unbiasedness, we have

$$\alpha = \frac{\beta \bar{d}}{N_v - 1 - \bar{d}} = \Theta\left(\frac{\log N_v}{N_v}\right).$$

(i)  $0 < \zeta \leq 2$

Note that

$$\alpha(N_v - 1) + \beta - (\alpha + \beta) \frac{\sum_{i=1}^{N_v} d_i^2}{\sum_{i=1}^{N_v} d_i} = -\beta\kappa + \alpha(N_v - \kappa - 1) + \beta,$$

and

$$\frac{\mathbb{E}X}{\mathbb{E}Y} = \Theta\left(\frac{\sum_{i=1}^{N_v} d_i^2}{\sum_{i=1}^{N_v} d_i}\right).$$

By Theorem A.2, for any  $\eta > 0$ , we have

$$\begin{aligned} \text{Bias}[\tilde{\kappa}] &= -\beta(2 - \alpha - \beta)\kappa + (2 - \alpha - \beta)[\alpha(N_v - \kappa - 1) + \beta] + \mathcal{O}\left(\frac{1}{(\mathbb{E}Y)^{1/(2+\eta)}} \frac{\mathbb{E}X}{\mathbb{E}Y}\right) \\ &= -\beta(2 - \alpha - \beta)\kappa + \mathcal{O}\left(\max\left\{\log N_v, \frac{\kappa}{(N_v \log N_v)^{1/(2+\eta)}}\right\}\right) \end{aligned}$$

as  $N_v \rightarrow \infty$ .

(ii)  $\zeta > 2$

Note that

$$\alpha(N_v - 1) + \beta - (\alpha + \beta) \frac{\sum_{i=1}^{N_v} d_i^2}{\sum_{i=1}^{N_v} d_i} \sim -\frac{\beta \bar{d}}{\zeta(\zeta - 2)} - \frac{\beta(\bar{d})^2}{\zeta(\zeta - 2)(N_v - 1 - \bar{d})} + \beta,$$

and

$$\frac{\mathbb{E}X}{\mathbb{E}Y} = \mathcal{O}(\bar{d}).$$

By Theorem A.2, for any  $\eta > 0$ , we have

$$\begin{aligned} \text{Bias}[\tilde{\kappa}] &= -(2 - \alpha - \beta) \frac{\beta \bar{d}}{\zeta(\zeta - 2)} + \mathcal{O}(1) \\ &= -\beta(2 - \alpha - \beta) \frac{\kappa}{(\zeta - 1)^2} + \mathcal{O}(1) \end{aligned}$$

as  $N_v \rightarrow \infty$ .

### Proof of Corollary A.3

By homogeneity, we obtain

$$\sum_{i=1}^{N_v} d_i^2 = (\bar{d} + 1) \bar{d} N_v.$$

By edge unbiasedness, we have

$$\alpha = \frac{\beta \bar{d}}{N_v - 1 - \bar{d}}. \tag{A.11}$$

Thus,

$$\alpha(N_v - 1) + \beta - (\alpha + \beta) \frac{\sum_{i=1}^{N_v} d_i^2}{\sum_{i=1}^{N_v} d_i} = -\alpha, \tag{A.12}$$

and

$$\frac{\mathbb{E}X}{\mathbb{E}Y} = \bar{d} + 1 - \alpha(2 - \alpha - \beta).$$

By Theorem A.2, for any  $\eta > 0$ , we have

$$\text{Bias}[\tilde{\kappa}] = \mathcal{O}\left(N_v^{c-\frac{c+1}{2+\eta}}\right) \text{ as } N_v \rightarrow \infty.$$

#### Proof of Corollary A.4

Note that the asymptotic notations for  $d_L$  and  $\sum_{i=1}^{N_v} d_i^2 / \sum_{i=1}^{N_v} d_i$  are same as equation (A.9) and equation (A.10).

By edge unbiasedness, we have

$$\alpha = \frac{\beta \bar{d}}{N_v - 1 - \bar{d}} = \Theta(N_v^{c-1}).$$

(i)  $0 < \zeta \leq 2$

Note that

$$\alpha(N_v - 1) + \beta - (\alpha + \beta) \frac{\sum_{i=1}^{N_v} d_i^2}{\sum_{i=1}^{N_v} d_i} = -\beta\kappa + \alpha(N_v - \kappa - 1) + \beta,$$

and

$$\frac{\mathbb{E}X}{\mathbb{E}Y} = \Theta\left(\frac{\sum_{i=1}^{N_v} d_i^2}{\sum_{i=1}^{N_v} d_i}\right).$$

By Theorem A.2, for any  $\eta > 0$ , we have

$$\begin{aligned} \text{Bias}[\tilde{\kappa}] &= -\beta(2 - \alpha - \beta)\kappa + (2 - \alpha - \beta)[\alpha(N_v - \kappa - 1) + \beta] + \mathcal{O}\left(\frac{1}{(\mathbb{E}Y)^{1/(2+\eta)}} \frac{\mathbb{E}X}{\mathbb{E}Y}\right) \\ &= -\beta(2 - \alpha - \beta)\kappa + \mathcal{O}\left(\max\left\{N_v^c, \frac{\kappa}{N_v^{(c+1)/(2+\eta)}}\right\}\right) \end{aligned}$$

as  $N_v \rightarrow \infty$ .

(ii)  $\zeta > 2$

Note that

$$\alpha(N_v - 1) + \beta - (\alpha + \beta) \frac{\sum_{i=1}^{N_v} d_i^2}{\sum_{i=1}^{N_v} d_i} \sim -\frac{\beta \bar{d}}{\zeta(\zeta - 2)} - \frac{\beta(\bar{d})^2}{\zeta(\zeta - 2)(N_v - 1 - \bar{d})} + \beta,$$

and

$$\frac{\mathbb{E}X}{\mathbb{E}Y} = \mathcal{O}(\bar{d}).$$

By Theorem A.2, for any  $\eta > 0$ , we have

$$\begin{aligned} \text{Bias}[\tilde{\kappa}] &= -(2 - \alpha - \beta) \frac{\beta \bar{d}}{\zeta(\zeta - 2)} + \mathcal{O}(\max\{N_v^{2c-1}, 1\}) \\ &= -\beta(2 - \alpha - \beta) \frac{\kappa}{(\zeta - 1)^2} + \mathcal{O}(\max\{N_v^{2c-1}, 1\}) \end{aligned}$$

as  $N_v \rightarrow \infty$ .

## A.5 Proofs of theorems for variances of the observed branching number

To show Theorem A.3 and Theorem A.4, we first introduce a useful lemma.

**Lemma A.1** *Under assumptions in Theorem A.3, for any  $\eta > 0$ , we have*

$$\begin{aligned} \mathbb{E} \left[ \left( \tilde{\kappa} - \frac{\mathbb{E}X}{\mathbb{E}Y} \right)^2 \right] &= \mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4} \right] + \mathcal{O} \left( \max \left\{ (\mathbb{E}Y)^{-1/(2+\eta)} \right. \right. \\ &\quad \left. \left. \cdot \mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4} \right], \mathbb{P}(Y = 0) \cdot \left[ \frac{\mathbb{E}X}{\mathbb{E}Y} \right]^2 \right\} \right) \end{aligned}$$

as  $N_v \rightarrow \infty$ .

**Proof.** Note that

$$\begin{aligned} & \mathbb{E} \left[ \left( \tilde{\kappa} - \frac{\mathbb{E}X}{\mathbb{E}Y} \right)^2 \right] - \mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4} \right] \\ &= \mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2 [(\mathbb{E}Y)^2 - Y^2]}{Y^2 (\mathbb{E}Y)^4} \cdot I_{\{Y>0\}} \right] + \left( \frac{\mathbb{E}X}{\mathbb{E}Y} \right)^2 \cdot \Pr(Y = 0). \end{aligned}$$

By triangle inequality and Jensen's inequality, we obtain

$$\begin{aligned} & \left| \mathbb{E} \left[ \left( \frac{X}{Y} \cdot I_{\{Y>0\}} - \frac{\mathbb{E}X}{\mathbb{E}Y} \right)^2 \right] - \mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4} \right] \right| \\ & \leq \frac{1}{(\mathbb{E}Y)^4} \mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2 |(\mathbb{E}Y)^2 - Y^2|}{Y^2} \cdot I_{\{Y>0\}} \right] + \left( \frac{\mathbb{E}X}{\mathbb{E}Y} \right)^2 \cdot \Pr(Y = 0). \quad (\text{A.13}) \end{aligned}$$

Next, we find an upper bound of

$$\mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2 |(\mathbb{E}Y)^2 - Y^2|}{Y^2} \cdot I_{\{Y>0\}} \right]. \quad (\text{A.14})$$

By additivity of expectation, for any  $\delta \in (0, 1)$ , (A.14) equals

$$\mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2 |(\mathbb{E}Y)^2 - Y^2|}{Y^2} \cdot I_{\{Y>0\}} \cdot I_{\{|Y - \mathbb{E}Y| \geq \delta \mathbb{E}Y\}} \right] \quad (\text{A.15})$$

$$+ \mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2 |(\mathbb{E}Y)^2 - Y^2|}{Y^2} \cdot I_{\{Y>0\}} \cdot I_{\{|Y - \mathbb{E}Y| < \delta \mathbb{E}Y\}} \right]. \quad (\text{A.16})$$

For the first term in (A.15), by definitions of  $X$  and  $Y$ , we obtain  $\left| \frac{X\mathbb{E}Y - Y\mathbb{E}X}{Y} \right| \cdot I_{\{Y>0\}} < N_v^3$  and  $|(\mathbb{E}Y)^2 - Y^2| < N_v^4$ . Thus, we have

$$\begin{aligned} & \mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2 |(\mathbb{E}Y)^2 - Y^2|}{Y^2} \cdot I_{\{Y>0\}} \cdot I_{\{|Y - \mathbb{E}Y| \geq \delta \mathbb{E}Y\}} \right] \\ & < N_v^{10} \cdot \Pr(|Y - \mathbb{E}Y| \geq \delta \mathbb{E}Y). \end{aligned}$$

Then, by Chernoff Bound, we obtain

$$\mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2 |(\mathbb{E}Y)^2 - Y^2|}{Y^2} \cdot I_{\{Y>0\}} \cdot I_{\{|Y - \mathbb{E}Y| \geq \delta \mathbb{E}Y\}} \right] < 2N_v^{10} \cdot \exp \left( - \frac{\delta^2 \cdot \mathbb{E}Y}{6} \right).$$

For the second term in (A.15), when  $|\mathbb{E}Y - Y| < \delta\mathbb{E}Y$ ,  $Y > (1 - \delta)\mathbb{E}Y$  and  $Y < (1 + \delta)\mathbb{E}Y$ . And notice  $|(\mathbb{E}Y)^2 - Y^2| = (\mathbb{E}Y + Y)|\mathbb{E}Y - Y|$ , we have

$$\begin{aligned} & \mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2|(\mathbb{E}Y)^2 - Y^2|}{Y^2} \cdot I_{\{Y>0\}} \cdot I_{\{|Y-\mathbb{E}Y|<\delta\mathbb{E}Y\}}\right] \\ & < \frac{\delta(2+\delta)}{(1-\delta)^2} \cdot \mathbb{E}\left[(X\mathbb{E}Y - Y\mathbb{E}X)^2\right]. \end{aligned}$$

By (A.13), we show

$$\begin{aligned} & \left| \mathbb{E}\left[\left(\tilde{\kappa} - \frac{\mathbb{E}X}{\mathbb{E}Y}\right)^2\right] - \mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] \right| \\ & \leq \frac{2N_v^{10}}{(\mathbb{E}Y)^4} \cdot \exp\left(-\frac{\delta^2 \cdot \mathbb{E}Y}{6}\right) + \frac{\delta(2+\delta)}{(1-\delta)^2} \cdot \mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] + \left(\frac{\mathbb{E}X}{\mathbb{E}Y}\right)^2 \cdot \Pr(Y=0). \end{aligned} \tag{A.17}$$

$L_1(N_v)$  and  $L_2(N_v)$  denote the first two terms on the right sides in (A.17). We choose  $\delta = (\mathbb{E}Y)^{-1/(2+\eta)}$ ,  $\eta > 0$ , such that

$$L_1(N_v) = o\left(\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right]\right) \text{ and } L_2(N_v) = o\left(\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right]\right)$$

as  $N_v \rightarrow \infty$ . Under the assumption  $\mathbb{E}Y = \Omega(N_v)$  as  $N_v \rightarrow \infty$ ,  $1 - \delta = \mathcal{O}(1)$ . By L'Hopital's rule, we have

$$L_1(N_v) = o(L_2(N_v)) \text{ as } N_v \rightarrow \infty.$$

These complete the proof. □

### Proof of Theorem A.3

By the definition of variance, we have

$$\text{Var}[\tilde{\kappa}] = \text{Var}\left[\tilde{\kappa} - \frac{\mathbb{E}X}{\mathbb{E}Y}\right] = \mathbb{E}\left[\left(\tilde{\kappa} - \frac{\mathbb{E}X}{\mathbb{E}Y}\right)^2\right] - \left[\mathbb{E}(\tilde{\kappa}) - \frac{\mathbb{E}X}{\mathbb{E}Y}\right]^2.$$

(i) By Lemma A.1, for any  $\eta > 0$ , we obtain

$$\begin{aligned}\text{Var}[\tilde{\kappa}] &= \mathcal{O}\left(\mathbb{E}\left[\left(\tilde{\kappa} - \frac{\mathbb{E}X}{\mathbb{E}Y}\right)^2\right]\right) \\ &= \mathcal{O}\left(\max\left\{\mathbb{E}\left[\frac{(XEY - YEX)^2}{(\mathbb{E}Y)^4}\right], \mathbb{P}(Y = 0) \cdot \left[\frac{\mathbb{E}X}{\mathbb{E}Y}\right]^2\right\}\right).\end{aligned}$$

(ii) By triangle inequality, we have

$$\begin{aligned}&\left|\text{Var}[\tilde{\kappa}] - \mathbb{E}\left[\frac{(XEY - YEX)^2}{(\mathbb{E}Y)^4}\right]\right| \\ &\leq \left|\mathbb{E}\left[\left(\tilde{\kappa} - \frac{\mathbb{E}X}{\mathbb{E}Y}\right)^2\right] - \mathbb{E}\left[\frac{(XEY - YEX)^2}{(\mathbb{E}Y)^4}\right]\right| + \left[\mathbb{E}(\tilde{\kappa}) - \frac{\mathbb{E}X}{\mathbb{E}Y}\right]^2.\end{aligned}$$

Apply Lemma A.1 and Theorem A.1 and the rest follows.

#### Proof of Theorem A.4

Under assumptions in Theorem A.4, we have

$$\mathbb{P}(Y = 0) \left[\frac{\mathbb{E}X}{\mathbb{E}Y}\right]^2 = o\left(\mathbb{E}\left[\frac{(XEY - YEX)^2}{(\mathbb{E}Y)^4}\right]\right),$$

and there exist  $\eta_0, \lambda_0 > 0$ , such that

$$\begin{aligned}&\mathbb{P}(Y = 0) \left[\frac{\mathbb{E}X}{\mathbb{E}Y}\right]^2 \\ &= o\left(\max\left\{\frac{1}{(\mathbb{E}Y)^{-1/(2+\eta_0)}} \mathbb{E}\left[\frac{(XEY - YEX)^2}{(\mathbb{E}Y)^4}\right], \frac{1}{(\mathbb{E}Y)^{-2/(2+\lambda_0)}} \left[\frac{\mathbb{E}X}{\mathbb{E}Y}\right]^2\right\}\right).\end{aligned}$$

Apply Theorem A.3 and the rest follows.

## A.6 Proofs of corollaries for variances of the observed branching number

To prove corollaries for variances of the observed branching number, we first compute  $\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right]$ . Note that

$$\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] = \left[\frac{\mathbb{E}X}{\mathbb{E}Y}\right]^2 \cdot \left[\frac{\text{Var}X}{(\mathbb{E}X)^2} - 2\frac{\text{Cov}(X, Y)}{\mathbb{E}X\mathbb{E}Y} + \frac{\text{Var}Y}{(\mathbb{E}Y)^2}\right].$$

Under Assumption 2.1, 2.2 and 2.4, we have

$$\mathbb{E}Y = \sum_{i=1}^{N_v} d_i,$$

$$\mathbb{E}X = (1 - \alpha - \beta)^2 \sum_{i=1}^{N_v} d_i^2 + (2 - \alpha - \beta)[\alpha(N_v - 1) + \beta] \sum_{i=1}^{N_v} d_i,$$

$$\text{Var}Y = 2\beta(2 - \alpha - \beta) \sum_{i=1}^{N_v} d_i,$$

$$\begin{aligned} \text{Cov}(X, Y) &= 4(\beta - \alpha)(1 - \alpha - \beta)^2 \sum_{i=1}^{N_v} d_i^2 + 2\left\{\beta[(1 - \alpha)(1 - 2\alpha) - (1 - \beta)(1 - 2\beta)]\right. \\ &\quad \left.+ 2\alpha(N_v - 1)[\beta(1 - \beta) + (1 - \alpha)^2]\right\} \sum_{i=1}^{N_v} d_i, \end{aligned}$$

$$\begin{aligned} \text{Var}X &= 4(\beta - \alpha)(1 - \alpha - \beta)^3 \sum_{i=1}^{N_v} d_i^3 \\ &\quad + 2(1 - \alpha - \beta)^2 [19\alpha^2 + 9\beta^2 - 6\alpha(1 + 3\beta) - 4\beta \\ &\quad + 2\alpha(1 + 4\beta - 5\alpha)N_v] \sum_{i=1}^{N_v} d_i^2 + \left\{(1 - \alpha - \beta) \left[ -38\alpha^3 \right. \right. \\ &\quad \left. \left. + 2\alpha^2(17 + 11\beta) + \beta(1 - 6\beta + 6\beta^2) - \alpha(5 + 14\beta^2) \right. \right. \\ &\quad \left. \left. + 4\alpha(1 + 11\alpha^2 + 2\beta^2 - \alpha(9 + 5\beta))N_v + 4\alpha^2(2 - 3\alpha + \beta)N_v^2 \right] \right. \\ &\quad \left. + (1 - \alpha)(\alpha + \beta) \left[ 1 - 12\alpha + 20\alpha^2 + 6(1 - 3\alpha)\alpha N_v + 4\alpha^2 N_v^2 \right] \right\} \end{aligned}$$

$$\begin{aligned}
& + 2(1 - \alpha - \beta) \left[ 3\alpha^2 - \beta(1 - \beta) - \alpha(1 + 2\beta) + \alpha(1 - 2\alpha + \beta)N_v \right] \\
& + 4(1 - \alpha - \beta) \left[ -8\alpha^3 + 3\alpha(1 - \beta)\beta + (1 - \beta)^2\beta + 4\alpha^2(1 + \beta) \right] \\
& - 8\alpha \left[ 5\alpha^3 + (1 - \beta)^2\beta - \alpha^2(8 - 3\beta) + \alpha(3 - 2\beta - \beta^2) \right] N_v \\
& + 4\alpha^2 \left[ 2 + 3\alpha^2 - \beta - \beta^2 - \alpha(5 - 2\beta) \right] N_v^2 \\
& + 2\alpha(1 - \alpha)(\alpha + \beta)(N_v - 2) \left[ 1 + 2\alpha(N_v - 2) \right] \\
& + \beta \left[ \alpha(\alpha - 3) - \beta(\beta - 3) \right] + 2\alpha(N_v - 1) \left[ \beta(1 - \beta) + (1 - \alpha)^2 \right] \left. \vphantom{\beta} \right\} \sum_{i=1}^{N_v} d_i \\
& + 4(1 - \alpha - \beta)^2 \left[ \alpha(1 - \alpha) \sum_{i=1}^{N_v} \sum_{j \neq i} d_i d_j I_{\{A_{ij}=0\}} \right. \\
& \left. + \beta(1 - \beta) \sum_{i=1}^{N_v} \sum_{j \neq i} d_i d_j I_{\{A_{ij}=1\}} \right].
\end{aligned}$$

### Proof of Corollary A.5

By edge unbiasedness, we have

$$\alpha = \frac{\beta \bar{d}}{N_v - 1 - \bar{d}} = \Theta\left(\frac{\log N_v}{N_v}\right). \tag{A.18}$$

By homogeneity, we obtain

$$\begin{aligned}
\sum_{i=1}^{N_v} d_i^2 &= \Theta(N_v \log^2(N_v)), \\
\sum_{i=1}^{N_v} d_i^3 &= \Theta(N_v \log^3(N_v)).
\end{aligned}$$

Besides, by Young's inequality, we have

$$\begin{aligned} \sum_{i=1}^{N_v} \sum_{j \neq i} d_i d_j I_{\{A_{ij}=1\}} &\leq \left( \sum_{i=1}^{N_v} \sum_{j \neq i} d_i^2 d_j^2 \right)^{1/2} \cdot \left( \sum_{i=1}^{N_v} \sum_{j \neq i} I_{\{A_{ij}=1\}} \right)^{1/2} = \mathcal{O}(N_v^{3/2} \log^{5/2} N_v), \\ \sum_{i=1}^{N_v} \sum_{j \neq i} d_i d_j I_{\{A_{ij}=0\}} &\leq \left( \sum_{i=1}^{N_v} \sum_{j \neq i} d_i^2 d_j^2 \right)^{1/2} \cdot \left( \sum_{i=1}^{N_v} \sum_{j \neq i} I_{\{A_{ij}=0\}} \right)^{1/2} = \mathcal{O}(N_v^2 \log^2 N_v). \end{aligned}$$

Thus, we obtain

$$\begin{aligned} \frac{\mathbb{E}X}{\mathbb{E}Y} &= \Theta(\log N_v), \\ \frac{\text{Var}X}{(\mathbb{E}X)^2} &= \mathcal{O}\left(\frac{1}{N_v^{1/2} \log^{3/2} N_v}\right), \\ \frac{\text{Cov}(X, Y)}{\mathbb{E}X \mathbb{E}Y} &= \Theta\left(\frac{1}{N_v \log N_v}\right), \\ \frac{\text{Var}Y}{(\mathbb{E}Y)^2} &= \Theta\left(\frac{1}{N_v \log N_v}\right). \end{aligned}$$

Then, we have

$$\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] = \mathcal{O}\left(\left(\frac{\log N_v}{N_v}\right)^{1/2}\right).$$

By Theorem A.4,

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}\left(\left(\frac{\log N_v}{N_v}\right)^{1/2}\right) \text{ as } N_v \rightarrow \infty.$$

### Proof of Corollary A.6

By edge unbiasedness, we have

$$\alpha = \frac{\beta \bar{d}}{N_v - 1 - \bar{d}} = \Theta\left(\frac{\log N_v}{N_v}\right). \quad (\text{A.19})$$

Next, we compute  $\sum_{i=1}^{N_v} d_i^3$  for different values of  $\zeta$ .

$$\int_{d_L}^{N_v-1} x^3 \cdot \frac{\zeta d_L^\zeta}{1 - \left(\frac{d_L}{N_v-1}\right)^\zeta} x^{-(\zeta+1)} dx = \begin{cases} \zeta d_L^\zeta \cdot \frac{\log\left(\frac{N_v-1}{d_L}\right)}{1 - \left(\frac{d_L}{N_v-1}\right)^\zeta}, & \text{if } \zeta = 3 \\ \frac{\zeta d_L^3}{\zeta - 3} \cdot \frac{1 - \left(\frac{d_L}{N_v-1}\right)^{\zeta-3}}{1 - \left(\frac{d_L}{N_v-1}\right)^\zeta}, & \text{otherwise.} \end{cases} \quad (\text{A.20})$$

Thus, we obtain

$$\sum_{i=1}^{N_v} d_i^3 = \begin{cases} \Theta(N_v^3 \log N_v), & \text{if } 0 < \zeta < 1 \\ \Theta(N_v^3), & \text{if } \zeta = 1 \\ \Theta(N_v^{4-\zeta} \cdot \log^\zeta N_v), & \text{if } 1 < \zeta < 3 \\ \Theta(N_v \log^4 N_v), & \text{if } \zeta = 3 \\ \Theta(N_v \log^3 N_v), & \text{if } \zeta > 3. \end{cases}$$

Equation (A.8) leads to

$$\sum_{i=1}^{N_v} d_i^2 = \begin{cases} \Theta(N_v^2 \log N_v), & \text{if } 0 < \zeta < 1 \\ \Theta(N_v^2), & \text{if } \zeta = 1 \\ \Theta(N_v^{3-\zeta} \cdot \log^\zeta N_v), & \text{if } 1 < \zeta < 2 \\ \Theta(N_v \log^3 N_v), & \text{if } \zeta = 2 \\ \Theta(N_v \log^2 N_v), & \text{if } \zeta > 2. \end{cases}$$

In addition, by Young's inequality, we have

$$\begin{aligned} \sum_{i=1}^{N_v} \sum_{j \neq i} d_i d_j I_{\{A_{ij}=1\}} &\leq \left( \sum_{i=1}^{N_v} \sum_{j \neq i} d_i^2 d_j^2 \right)^{1/2} \cdot \left( \sum_{i=1}^{N_v} \sum_{j \neq i} I_{\{A_{ij}=1\}} \right)^{1/2}, \\ \sum_{i=1}^{N_v} \sum_{j \neq i} d_i d_j I_{\{A_{ij}=0\}} &\leq \left( \sum_{i=1}^{N_v} \sum_{j \neq i} d_i^2 d_j^2 \right)^{1/2} \cdot \left( \sum_{i=1}^{N_v} \sum_{j \neq i} I_{\{A_{ij}=0\}} \right)^{1/2}. \end{aligned} \quad (\text{A.21})$$

Thus, we obtain

$$\sum_{i=1}^{N_v} \sum_{j \neq i} d_i d_j I_{\{A_{ij}=1\}} = \begin{cases} \mathcal{O}(N_v^{5/2} \log^{3/2} N_v), & \text{if } 0 < \zeta < 1 \\ \mathcal{O}(N_v^{5/2} \log^{1/2} N_v), & \text{if } \zeta = 1 \\ \mathcal{O}(N_v^{7/2-\zeta} \cdot \log^{\zeta+1/2} N_v), & \text{if } 1 < \zeta < 2 \\ \mathcal{O}(N_v^{3/2} \log^{7/2} N_v), & \text{if } \zeta = 2 \\ \mathcal{O}(N_v^{3/2} \log^{5/2} N_v), & \text{if } \zeta > 2, \end{cases}$$

and

$$\sum_{i=1}^{N_v} \sum_{j \neq i} d_i d_j I_{\{A_{ij}=0\}} = \begin{cases} \mathcal{O}(N_v^3 \log N_v), & \text{if } 0 < \zeta < 1 \\ \mathcal{O}(N_v^3), & \text{if } \zeta = 1 \\ \mathcal{O}(N_v^{4-\zeta} \cdot \log^\zeta N_v), & \text{if } 1 < \zeta < 2 \\ \mathcal{O}(N_v^2 \log^3 N_v), & \text{if } \zeta = 2 \\ \mathcal{O}(N_v^2 \log^2 N_v), & \text{if } \zeta > 2. \end{cases}$$

(i)  $0 < \zeta < 1$

Note that

$$\begin{aligned}\frac{\mathbb{E}X}{\mathbb{E}Y} &= \Theta(N_v), \\ \frac{\text{Var}X}{(\mathbb{E}X)^2} &= \Theta\left(\frac{1}{N_v \log N_v}\right), \\ \frac{\text{Cov}(X, Y)}{\mathbb{E}X\mathbb{E}Y} &= \Theta\left(\frac{1}{N_v \log N_v}\right), \\ \frac{\text{Var}Y}{(\mathbb{E}Y)^2} &= \Theta\left(\frac{1}{N_v \log N_v}\right).\end{aligned}$$

Then, we have

$$\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] = \mathcal{O}\left(\frac{N_v}{\log N_v}\right).$$

By Theorem A.4,

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}\left(\frac{N_v}{\log N_v}\right) \text{ as } N_v \rightarrow \infty.$$

(ii)  $\zeta = 1$

Note that

$$\begin{aligned}\frac{\mathbb{E}X}{\mathbb{E}Y} &= \Theta\left(\frac{N_v}{\log N_v}\right), \\ \frac{\text{Var}X}{(\mathbb{E}X)^2} &= \Theta\left(\frac{1}{N_v}\right), \\ \frac{\text{Cov}(X, Y)}{\mathbb{E}X\mathbb{E}Y} &= \Theta\left(\frac{1}{N_v \log N_v}\right), \\ \frac{\text{Var}Y}{(\mathbb{E}Y)^2} &= \Theta\left(\frac{1}{N_v \log N_v}\right).\end{aligned}$$

Then, we have

$$\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] = \mathcal{O}\left(\frac{N_v}{\log^2 N_v}\right).$$

By Theorem A.4,

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}\left(\frac{N_v}{\log^2 N_v}\right) \text{ as } N_v \rightarrow \infty.$$

(iii)  $1 < \zeta < 2$

Note that

$$\begin{aligned} \frac{\mathbb{E}X}{\mathbb{E}Y} &= \Theta\left(N_v^{2-\zeta} \cdot \log^{\zeta-1} N_v\right), \\ \frac{\text{Var}X}{(\mathbb{E}X)^2} &= \Theta\left(\frac{1}{N_v^{2-\zeta} \cdot \log^\zeta N_v}\right), \\ \frac{\text{Cov}(X, Y)}{\mathbb{E}X\mathbb{E}Y} &= \Theta\left(\frac{1}{N_v \log N_v}\right), \\ \frac{\text{Var}Y}{(\mathbb{E}Y)^2} &= \Theta\left(\frac{1}{N_v \log N_v}\right). \end{aligned}$$

Then, we have

$$\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] = \mathcal{O}\left(\left(\frac{N_v}{\log N_v}\right)^{2-\zeta}\right).$$

By Theorem A.4,

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}\left(\left(\frac{N_v}{\log N_v}\right)^{2-\zeta}\right) \text{ as } N_v \rightarrow \infty.$$

(iv)  $\zeta = 2$

Note that

$$\begin{aligned} \frac{\mathbb{E}X}{\mathbb{E}Y} &= \Theta\left(\log^2 N_v\right), \\ \frac{\text{Var}X}{(\mathbb{E}X)^2} &= \Theta\left(\frac{1}{\log^4 N_v}\right), \\ \frac{\text{Cov}(X, Y)}{\mathbb{E}X\mathbb{E}Y} &= \Theta\left(\frac{1}{N_v \log N_v}\right), \\ \frac{\text{Var}Y}{(\mathbb{E}Y)^2} &= \Theta\left(\frac{1}{N_v \log N_v}\right). \end{aligned}$$

Then, we have

$$\mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4} \right] = \mathcal{O}(1).$$

By Theorem A.4,

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}(1) \text{ as } N_v \rightarrow \infty.$$

(v)  $2 < \zeta < 5/2$

Note that

$$\begin{aligned} \frac{\mathbb{E}X}{\mathbb{E}Y} &= \Theta(\log N_v), \\ \frac{\text{Var}X}{(\mathbb{E}X)^2} &= \Theta\left(\frac{1}{N_v^{\zeta-2} \log^{4-\zeta} N_v}\right), \\ \frac{\text{Cov}(X, Y)}{\mathbb{E}X\mathbb{E}Y} &= \Theta\left(\frac{1}{N_v \log N_v}\right), \\ \frac{\text{Var}Y}{(\mathbb{E}Y)^2} &= \Theta\left(\frac{1}{N_v \log N_v}\right). \end{aligned}$$

Then, we have

$$\mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4} \right] = \mathcal{O}\left(\left(\frac{\log N_v}{N_v}\right)^{\zeta-2}\right).$$

By Theorem A.4,

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}\left(\left(\frac{\log N_v}{N_v}\right)^{\zeta-2}\right) \text{ as } N_v \rightarrow \infty.$$

(vi)  $\zeta \geq 5/2$

Note that

$$\begin{aligned}\frac{\mathbb{E}X}{\mathbb{E}Y} &= \Theta(\log N_v), \\ \frac{\text{Var}X}{(\mathbb{E}X)^2} &= \mathcal{O}\left(\frac{1}{N_v^{1/2} \log^{3/2} N_v}\right), \\ \frac{\text{Cov}(X, Y)}{\mathbb{E}X\mathbb{E}Y} &= \Theta\left(\frac{1}{N_v \log N_v}\right), \\ \frac{\text{Var}Y}{(\mathbb{E}Y)^2} &= \Theta\left(\frac{1}{N_v \log N_v}\right).\end{aligned}$$

Then, we have

$$\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] = \mathcal{O}\left(\left(\frac{\log N_v}{N_v}\right)^{1/2}\right).$$

By Theorem A.4,

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}\left(\left(\frac{\log N_v}{N_v}\right)^{1/2}\right) \text{ as } N_v \rightarrow \infty.$$

### Proof of Corollary A.7

By edge unbiasedness, we have

$$\alpha = \frac{\beta \bar{d}}{N_v - 1 - \bar{d}} = \Theta(N_v^{c-1}). \tag{A.22}$$

By homogeneity, we obtain

$$\begin{aligned}\sum_{i=1}^{N_v} d_i^2 &= \Theta(N_v^{2c+1}), \\ \sum_{i=1}^{N_v} d_i^3 &= \Theta(N_v^{3c+1}).\end{aligned}$$

Besides, by Young's inequality, we have

$$\begin{aligned} \sum_{i=1}^{N_v} \sum_{j \neq i} d_i d_j I_{\{A_{ij}=1\}} &\leq \left( \sum_{i=1}^{N_v} \sum_{j \neq i} d_i^2 d_j^2 \right)^{1/2} \cdot \left( \sum_{i=1}^{N_v} \sum_{j \neq i} I_{\{A_{ij}=1\}} \right)^{1/2} = \mathcal{O}(N_v^{(5c+3)/2}), \\ \sum_{i=1}^{N_v} \sum_{j \neq i} d_i d_j I_{\{A_{ij}=0\}} &\leq \left( \sum_{i=1}^{N_v} \sum_{j \neq i} d_i^2 d_j^2 \right)^{1/2} \cdot \left( \sum_{i=1}^{N_v} \sum_{j \neq i} I_{\{A_{ij}=0\}} \right)^{1/2} = \mathcal{O}(N_v^{2c+2}). \end{aligned}$$

Thus, we obtain

$$\begin{aligned} \frac{\mathbb{E}X}{\mathbb{E}Y} &= \Theta(N_v^c), \\ \frac{\text{Var}X}{(\mathbb{E}X)^2} &= \mathcal{O}\left(\frac{1}{N_v^{(3c+1)/2}}\right), \\ \frac{\text{Cov}(X, Y)}{\mathbb{E}X\mathbb{E}Y} &= \Theta\left(\frac{1}{N_v^{c+1}}\right), \\ \frac{\text{Var}Y}{(\mathbb{E}Y)^2} &= \Theta\left(\frac{1}{N_v^{c+1}}\right). \end{aligned}$$

Then, we have

$$\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] = \mathcal{O}(N_v^{(c-1)/2}).$$

By Theorem A.4,

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}(N_v^{(c-1)/2}) \text{ as } N_v \rightarrow \infty.$$

### Proof of Corollary A.8

By edge unbiasedness, we have

$$\alpha = \frac{\beta \bar{d}}{N_v - 1 - \bar{d}} = \Theta(N_v^{c-1}). \quad (\text{A.23})$$

By equation (A.20), we have

$$\sum_{i=1}^{N_v} d_i^3 = \begin{cases} \Theta(N_v^{c+3}), & \text{if } 0 < \zeta < 1 \\ \Theta(N_v^{c+3}/\log N_v), & \text{if } \zeta = 1 \\ \Theta(N_v^{4-\zeta+c\zeta}), & \text{if } 1 < \zeta < 3 \\ \Theta(N_v^{3c+1} \cdot \log N_v), & \text{if } \zeta = 3 \\ \Theta(N_v^{3c+1}), & \text{if } \zeta > 3. \end{cases}$$

Equation (A.8) leads to

$$\sum_{i=1}^{N_v} d_i^2 = \begin{cases} \Theta(N_v^{c+2}), & \text{if } 0 < \zeta < 1 \\ \Theta(N_v^{c+2}/\log N_v), & \text{if } \zeta = 1 \\ \Theta(N_v^{3-\zeta+c\zeta}), & \text{if } 1 < \zeta < 2 \\ \Theta(N_v^{2c+1} \cdot \log N_v), & \text{if } \zeta = 2 \\ \Theta(N_v^{2c+1}), & \text{if } \zeta > 2. \end{cases}$$

In addition, by equation (A.21), we obtain

$$\sum_{i=1}^{N_v} \sum_{j \neq i} d_i d_j I_{\{A_{ij}=1\}} = \begin{cases} \mathcal{O}(N_v^{(3c+5)/2}), & \text{if } 0 < \zeta < 1 \\ \mathcal{O}(N_v^{(3c+5)/2}/\log N_v), & \text{if } \zeta = 1 \\ \mathcal{O}(N_v^{7/2-\zeta+c(\zeta+1/2)}), & \text{if } 1 < \zeta < 2 \\ \mathcal{O}(N_v^{(5c+3)/2} \cdot \log N_v), & \text{if } \zeta = 2 \\ \mathcal{O}(N_v^{(5c+3)/2}), & \text{if } \zeta > 2, \end{cases}$$

and

$$\sum_{i=1}^{N_v} \sum_{j \neq i} d_i d_j I_{\{A_{ij}=0\}} = \begin{cases} \mathcal{O}(N_v^{c+3}), & \text{if } 0 < \zeta < 1 \\ \mathcal{O}(N_v^{c+3}/\log N_v), & \text{if } \zeta = 1 \\ \mathcal{O}(N_v^{4-\zeta+c\zeta}), & \text{if } 1 < \zeta < 2 \\ \mathcal{O}(N_v^{2c+2} \cdot \log N_v), & \text{if } \zeta = 2 \\ \mathcal{O}(N_v^{2c+2}), & \text{if } \zeta > 2. \end{cases}$$

(i)  $0 < \zeta < 1$

Note that

$$\begin{aligned} \frac{\mathbb{E}X}{\mathbb{E}Y} &= \Theta(N_v), \\ \frac{\text{Var}X}{(\mathbb{E}X)^2} &= \Theta\left(\frac{1}{N_v^{c+1}}\right), \\ \frac{\text{Cov}(X, Y)}{\mathbb{E}X\mathbb{E}Y} &= \Theta\left(\frac{1}{N_v^{c+1}}\right), \\ \frac{\text{Var}Y}{(\mathbb{E}Y)^2} &= \Theta\left(\frac{1}{N_v^{c+1}}\right). \end{aligned}$$

Then, we have

$$\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] = \mathcal{O}(N_v^{1-c}).$$

By Theorem A.4,

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}(N_v^{1-c}) \text{ as } N_v \rightarrow \infty.$$

(ii)  $\zeta = 1$

Note that

$$\begin{aligned}\frac{\mathbb{E}X}{\mathbb{E}Y} &= \Theta\left(\frac{N_v}{\log N_v}\right), \\ \frac{\text{Var}X}{(\mathbb{E}X)^2} &= \Theta\left(\frac{\log N_v}{N_v^{c+1}}\right), \\ \frac{\text{Cov}(X, Y)}{\mathbb{E}X\mathbb{E}Y} &= \Theta\left(\frac{1}{N_v^{c+1}}\right), \\ \frac{\text{Var}Y}{(\mathbb{E}Y)^2} &= \Theta\left(\frac{1}{N_v^{c+1}}\right).\end{aligned}$$

Then, we have

$$\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] = \mathcal{O}\left(\frac{N_v^{1-c}}{\log N_v}\right).$$

By Theorem A.4,

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}\left(\frac{N_v^{1-c}}{\log N_v}\right) \text{ as } N_v \rightarrow \infty.$$

(iii)  $1 < \zeta < 2$

Note that

$$\begin{aligned}\frac{\mathbb{E}X}{\mathbb{E}Y} &= \Theta\left(N_v^{2-\zeta+c(\zeta-1)}\right), \\ \frac{\text{Var}X}{(\mathbb{E}X)^2} &= \Theta\left(\frac{1}{N_v^{2-\zeta+c\zeta}}\right), \\ \frac{\text{Cov}(X, Y)}{\mathbb{E}X\mathbb{E}Y} &= \Theta\left(\frac{1}{N_v^{c+1}}\right), \\ \frac{\text{Var}Y}{(\mathbb{E}Y)^2} &= \Theta\left(\frac{1}{N_v^{c+1}}\right).\end{aligned}$$

Then, we have

$$\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] = \mathcal{O}\left(N_v^{(2-\zeta)(1-c)}\right).$$

By Theorem A.4,

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}\left(N_v^{(2-\zeta)(1-c)}\right) \text{ as } N_v \rightarrow \infty.$$

(iv)  $\zeta = 2$

Note that

$$\begin{aligned} \frac{\mathbb{E}X}{\mathbb{E}Y} &= \Theta\left(N_v^c \cdot \log N_v\right), \\ \frac{\text{Var}X}{(\mathbb{E}X)^2} &= \Theta\left(\frac{1}{N_v^{2c} \cdot \log^2 N_v}\right), \\ \frac{\text{Cov}(X, Y)}{\mathbb{E}X\mathbb{E}Y} &= \Theta\left(\frac{1}{N_v^{c+1}}\right), \\ \frac{\text{Var}Y}{(\mathbb{E}Y)^2} &= \Theta\left(\frac{1}{N_v^{c+1}}\right). \end{aligned}$$

Then, we have

$$\mathbb{E}\left[\frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4}\right] = \mathcal{O}(1).$$

By Theorem A.4,

$$\text{Var}[\tilde{\kappa}] = \mathcal{O}(1) \text{ as } N_v \rightarrow \infty.$$

(v)  $2 < \zeta < 5/2$

Note that

$$\begin{aligned} \frac{\mathbb{E}X}{\mathbb{E}Y} &= \Theta\left(N_v^c\right), \\ \frac{\text{Var}X}{(\mathbb{E}X)^2} &= \Theta\left(\frac{1}{N_v^{\zeta-2+c(4-\zeta)}}\right), \\ \frac{\text{Cov}(X, Y)}{\mathbb{E}X\mathbb{E}Y} &= \Theta\left(\frac{1}{N_v^{c+1}}\right), \\ \frac{\text{Var}Y}{(\mathbb{E}Y)^2} &= \Theta\left(\frac{1}{N_v^{c+1}}\right). \end{aligned}$$

Then, we have

$$\mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4} \right] = \mathcal{O} \left( N_v^{(2-\zeta)(1-c)} \right).$$

By Theorem A.4,

$$\text{Var}[\tilde{\kappa}] = \mathcal{O} \left( N_v^{(2-\zeta)(1-c)} \right) \text{ as } N_v \rightarrow \infty.$$

(vi)  $\zeta \geq 5/2$

Note that

$$\begin{aligned} \frac{\mathbb{E}X}{\mathbb{E}Y} &= \Theta \left( N_v^c \right), \\ \frac{\text{Var}X}{(\mathbb{E}X)^2} &= \mathcal{O} \left( \frac{1}{N_v^{(3c+1)/2}} \right), \\ \frac{\text{Cov}(X, Y)}{\mathbb{E}X\mathbb{E}Y} &= \Theta \left( \frac{1}{N_v^{c+1}} \right), \\ \frac{\text{Var}Y}{(\mathbb{E}Y)^2} &= \Theta \left( \frac{1}{N_v^{c+1}} \right). \end{aligned}$$

Then, we have

$$\mathbb{E} \left[ \frac{(X\mathbb{E}Y - Y\mathbb{E}X)^2}{(\mathbb{E}Y)^4} \right] = \mathcal{O} \left( N_v^{(c-1)/2} \right).$$

By Theorem A.4,

$$\text{Var}[\tilde{\kappa}] = \mathcal{O} \left( N_v^{(c-1)/2} \right) \text{ as } N_v \rightarrow \infty.$$

## A.7 Proofs of theorem for the method-of-moments estimator

Chang et al. (2020) provide joint inference of higher-order subgraph densities with unknown error rates. Mimicking their proofs, we can easily obtain the asymptotic joint normal distribution of  $\hat{C}_{\mathcal{V}_1}$  and  $\hat{C}_{\mathcal{V}_2}$ . Then, by the delta method, we can derive the asymptotic normal distribution of  $\hat{\kappa}$ .

## A.8 Algorithm for estimation of asymptotic variance of method-of-moments estimator

To evaluate the asymptotic variance of method-of-moments estimator  $\hat{\kappa}$ , we first estimate the asymptotic variance of  $(\hat{C}_{\mathcal{V}_1}, \hat{C}_{\mathcal{V}_2})$  by the method in Section 4 of Chang et al. (2020). Then, we use the delta method to obtain the estimation of the asymptotic variance of  $\hat{\kappa}$ . The detail is shown in Algorithm A.1.

$$\begin{aligned}
\hat{\Sigma} &= (\hat{\sigma}_{ij})_{3 \times 3}, \\
\hat{\Delta} &= \hat{k}_3^{-1} \cdot \begin{pmatrix} \hat{C}_{\mathcal{V}_1} - 1 & \hat{C}_{\mathcal{V}_1} \\ 2\hat{C}_{\mathcal{V}_2} - 2\hat{C}_{\mathcal{V}_1} & 2\hat{C}_{\mathcal{V}_2} \end{pmatrix}, \\
\hat{\mathbf{G}} &= \hat{k}_3^{-2} \cdot \begin{pmatrix} (1 - \hat{\delta})^{-1}\{(1 - 2\hat{\beta})\hat{\alpha} + \hat{\beta}^2\} & (1 - \hat{\delta})^{-1}(\hat{\alpha} - 2\hat{\beta}) & (1 - \hat{\delta})^{-1} \\ -\hat{\delta}^{-1}\{(1 - 2\hat{\alpha})\hat{\beta} + \hat{\alpha}^2\} & \hat{\delta}^{-1}(\hat{\beta} - 2\hat{\alpha} + 1) & -\hat{\delta}^{-1} \end{pmatrix}, \\
\hat{\mathbf{H}} &= \frac{1}{3} \cdot \begin{pmatrix} \hat{C}_{\mathcal{V}_1} & \hat{k}_3^{-1} \\ 2\hat{C}_{\mathcal{V}_2} & 2\hat{k}_3^{-1}\hat{C}_{\mathcal{V}_1} \end{pmatrix} \\
&\quad \cdot \begin{pmatrix} 6\hat{k}_4 & 3(\hat{k}_4^2 - \hat{k}_1 - \hat{k}_2) & 2\{\hat{k}_4(-6\hat{\alpha}\hat{\beta} + 3\hat{k}_3^2 - 4\hat{k}_3) + (1 - \hat{\alpha})(\hat{\beta} - 2\hat{\alpha})\} \\ 6\hat{k}_1 & 3\hat{k}_1(1 - 2\hat{\alpha}) & 2\hat{k}_1(1 - \hat{\alpha})(1 - 3\hat{\alpha}) \end{pmatrix},
\end{aligned} \tag{A.24}$$

where  $\hat{k}_1 = \hat{\alpha}(1 - \hat{\alpha})$ ,  $\hat{k}_2 = \hat{\beta}(1 - \hat{\beta})$ ,  $\hat{k}_3 = 1 - \hat{\alpha} - \hat{\beta}$ ,  $\hat{k}_4 = \hat{\beta} - \hat{\alpha}$ ,  $\hat{\sigma}_{11} = \hat{\delta}\hat{k}_2 + (1 - \hat{\delta})\hat{k}_1$ ,  $\hat{\sigma}_{22} = \hat{\delta}\hat{k}_2(1/2 - \hat{k}_2) + (1 - \hat{\delta})\hat{k}_1(1/2 - \hat{k}_1)$ ,  $\hat{\sigma}_{33} = \hat{\delta}\hat{\beta}\hat{k}_2(1/3 - \hat{\beta}\hat{k}_2) + (1 - \hat{\delta})\hat{k}_1(1 - \hat{\alpha})\{1/3 - \hat{k}_1(1 - \hat{\alpha})\}$ ,  $\hat{\sigma}_{12} = \hat{\sigma}_{21} = \hat{\delta}\hat{k}_2(\hat{\beta} - 1/2) + (1 - \hat{\delta})\hat{k}_1(1/2 - \hat{\alpha})$ ,  $\hat{\sigma}_{13} = \hat{\sigma}_{31} = \hat{\delta}\hat{k}_2(\hat{\beta}^2/3 - 2\hat{k}_2/3) + (1 - \hat{\delta})\hat{k}_1\{(1 - \hat{\alpha})^2/3 - 2\hat{k}_1/3\}$ , and  $\hat{\sigma}_{23} = \hat{\sigma}_{32} = \hat{\delta}\hat{\beta}\hat{k}_2(1/3 - \hat{k}_2) + (1 - \hat{\delta})(1 - \hat{\alpha})\hat{k}_1(1/3 - \hat{k}_1)$ .

---

**Algorithm A.1** Estimation of asymptotic variance of method-of-moments estimator

---

$\hat{\kappa}$

---

**Input:**  $\tilde{\mathbf{A}} = (\tilde{A}_{i,j})_{N_v \times N_v}$ ,  $\varepsilon$ ,  $N_b$ ,  $\hat{\alpha}$ ,  $\hat{\beta}$ ,  $\hat{k}_3$ ,  $\hat{C}_{\mathcal{V}_1}$ ,  $\hat{C}_{\mathcal{V}_2}$ ,  $\hat{\delta}$

**Output:**  $\widehat{\text{Var}}(\hat{\kappa})$

**if**  $|\hat{\alpha} - \hat{\beta}| < \varepsilon$  **then**

$\xi_2 = \hat{\alpha}$ ,  $\xi_1 = 1 - 2\xi_2$ ;

**if**  $\hat{\beta} - \hat{\alpha} > \varepsilon$  **then**

$t_1 = \sqrt{1 - 4\hat{\alpha}(1 - \hat{\beta})}$ ,  $t_2 = \sqrt{1 - 4\hat{\beta}(1 - \hat{\alpha})}$ ,  $\xi_2 = (1 - t_1)/2$ ;

**if**  $t_1 + t_2 < 0.5$  **then**

$\xi_1 = (t_1 + t_2)/2$ ;

**else**

$\xi_1 = (t_1 - t_2)/2$ ;

**if**  $\hat{\alpha} - \hat{\beta} > \varepsilon$  **then**

$t_1 = \sqrt{1 - 4\hat{\alpha}(1 - \hat{\beta})}$ ,  $t_2 = \sqrt{1 - 4\hat{\beta}(1 - \hat{\alpha})}$ ,  $\xi_2 = (1 + t_1)/2$ ,  $\xi_1 = (t_2 - t_1)/2$ ;

**for**  $n_b = 1 : N_b$  **do**

**for**  $i = 1 : N_v$  **do**

**for**  $j = i + 1 : N_v$  **do**

Draw  $\eta_{i,j}$  from distribution  $\mathbb{P}(\eta_{i,j} = 0) = \xi_1$ ,  $\mathbb{P}(\eta_{i,j} = 1) = \xi_2$  and  $\mathbb{P}(\eta_{i,j} = -1) = 1 - \xi_1 - \xi_2$ ;

Compute  $\tilde{A}_{i,j}^\dagger = \tilde{A}_{i,j}I(\eta_{i,j} = 0) + I(\eta_{i,j} = 1)$ ;

Compute  $\hat{\tilde{A}}_{i,j}^\dagger = \hat{\tilde{A}}_{j,i}^\dagger = \tilde{A}_{i,j}^\dagger - \tilde{A}_{i,j}\xi_1 - \xi_2$ ;

$\hat{S}_{\mathcal{V}_1, n_b}^\dagger \leftarrow \frac{1}{\hat{k}_3} \sqrt{\frac{2}{N_v(N_v-1)}} \sum_{i < j} \hat{\tilde{A}}_{i,j}^\dagger$ ;

$\hat{S}_{\mathcal{V}_2, n_b}^\dagger \leftarrow \frac{1}{\hat{k}_3^2(N_v-2)} \sqrt{\frac{1}{2N_v(N_v-1)}} \sum_{i \neq j \neq l} \{\hat{\tilde{A}}_{i,j}^\dagger (\tilde{A}_{j,l} - \hat{\alpha}) + \hat{\tilde{A}}_{j,l}^\dagger (\tilde{A}_{i,j} - \hat{\alpha})\}$ ;

Compute  $\hat{\mathbf{V}}_{N_v} = \hat{\mathbf{V}}_{1, N_v} + \hat{\mathbf{V}}_{2, N_v} + \hat{\mathbf{V}}_{3, N_v}$ ,

where  $\hat{\mathbf{V}}_{1, N_v} = \begin{bmatrix} \text{Var}(\hat{S}_{\mathcal{V}_1}^\dagger) & \text{Cov}(\hat{S}_{\mathcal{V}_1}^\dagger, \hat{S}_{\mathcal{V}_2}^\dagger) \\ \text{Cov}(\hat{S}_{\mathcal{V}_1}^\dagger, \hat{S}_{\mathcal{V}_2}^\dagger) & \text{Var}(\hat{S}_{\mathcal{V}_2}^\dagger) \end{bmatrix}$ ,  $\hat{\mathbf{V}}_{2, N_v} = \hat{\mathbf{\Delta}} \hat{\mathbf{G}} \hat{\mathbf{\Sigma}} \hat{\mathbf{G}}^\top \hat{\mathbf{\Delta}}^\top$ ,

$\hat{\mathbf{V}}_{3, N_v} = (\hat{\mathbf{H}} \hat{\mathbf{G}}^\top \hat{\mathbf{\Delta}}^\top + \hat{\mathbf{\Delta}} \hat{\mathbf{G}} \hat{\mathbf{H}}^\top)/2$ ,  $\hat{\mathbf{\Delta}}$ ,  $\hat{\mathbf{G}}$ ,  $\hat{\mathbf{\Sigma}}$ , and  $\hat{\mathbf{H}}$  defined in (A.24);

Compute  $\widehat{\text{Var}}(\hat{\kappa}) = (N_v - 2)^2 [-\hat{C}_{\mathcal{V}_2}/\hat{C}_{\mathcal{V}_1}^2, 1/\hat{C}_{\mathcal{V}_1}] \hat{\mathbf{V}}_{N_v} [-\hat{C}_{\mathcal{V}_2}/\hat{C}_{\mathcal{V}_1}^2, 1/\hat{C}_{\mathcal{V}_1}]^\top$ .

---

## Appendix B

# Supplementary Materials to Chapter 3

We provide proofs of all propositions and theorems presented in Chapter 3.

### B.1 Proofs of propositions for noise-free networks

#### Proof of Proposition 3.1

Notice that

$$\begin{aligned} \text{Var}\left[\hat{y}(c_k)\right] &= \frac{1}{N_v^2} \left\{ \sum_{i=1}^{N_v} p_i^e(c_k) [1 - p_i^e(c_k)] \left[ \frac{y_i(c_k)}{p_i^e(c_k)} \right]^2 \right. \\ &\quad \left. + \sum_{i=1}^{N_v} \sum_{j \neq i} [p_{ij}^e(c_k) - p_i^e(c_k)p_j^e(c_k)] \frac{y_i(c_k)}{p_i^e(c_k)} \frac{y_j(c_k)}{p_j^e(c_k)} \right\}. \end{aligned}$$

Thus, under Condition 3.1 and 3.2, we have  $\text{Var}[\hat{y}(c_k)] = o(1)$  and  $\text{Var}[\hat{y}(c_l)] = o(1)$ .

Then, by Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \text{Var}\left[\hat{\tau}(c_k, c_l)\right] &= \text{Var}\left[\hat{y}(c_k)\right] + \text{Var}\left[\hat{y}(c_l)\right] - 2 \text{Cov}\left[\hat{y}(c_k), \hat{y}(c_l)\right] \\ &\leq \text{Var}\left[\hat{y}(c_k)\right] + \text{Var}\left[\hat{y}(c_l)\right] + 2\sqrt{\text{Var}\left[\hat{y}(c_k)\right]\text{Var}\left[\hat{y}(c_l)\right]}. \end{aligned}$$

Thus, we obtain  $\text{Var}[\hat{\tau}(c_k, c_l)] = o(1)$ . Since  $\mathbb{E}[\hat{\tau}(c_k, c_l)] = \tau(c_k, c_l)$ , we have  $\hat{\tau}(c_k, c_l) \xrightarrow{L_2} \tau(c_k, c_l)$ . This implies  $\hat{\tau}(c_k, c_l) \xrightarrow{P} \tau(c_k, c_l)$  as  $N_v \rightarrow \infty$ .

### Proof of Proposition 3.2

By Proposition 3.1, it suffices to show Condition 3.1 and 3.2 are satisfied. For notational simplicity, we define  $\mathbb{E}[1/p^e(c_k)] := \frac{1}{N_v} \sum_{i=1}^{N_v} 1/p_i^e(c_k)$ , i.e., the average of the inverse of the exposure probability over a finite population  $N_v$ .

Under Assumption 3.4, we obtain

$$\begin{aligned}\mathbb{E}[1/p^e(c_{10})] &= \frac{e^{\bar{d}/(1-p)} - 1}{(e^{\bar{d}} - 1)p} = \Theta(1/p), \\ \mathbb{E}[1/p^e(c_{00})] &= \frac{e^{\bar{d}/(1-p)} - 1}{(e^{\bar{d}} - 1)(1-p)} = \mathcal{O}(1).\end{aligned}$$

Note that  $\text{Var}[1 - (1-p)^d] = o(1)$ , where  $d$  is the degree variable. Thus,  $1 - (1-p)^d \xrightarrow{L_2} \lim_{N_v \rightarrow \infty} \mathbb{E}[1 - (1-p)^d]$ . This implies  $1 - (1-p)^d \xrightarrow{P} \lim_{N_v \rightarrow \infty} \mathbb{E}[1 - (1-p)^d]$ . By the continuous mapping theorem, we have  $\frac{p}{1 - (1-p)^d} \xrightarrow{P} \lim_{N_v \rightarrow \infty} \frac{p}{\mathbb{E}[1 - (1-p)^d]}$ . Thus, we obtain

$$\lim_{N_v \rightarrow \infty} \mathbb{E}\left[\frac{p}{1 - (1-p)^d}\right] = \lim_{N_v \rightarrow \infty} \frac{p}{\mathbb{E}[1 - (1-p)^d]}.$$

Since  $\mathbb{E}[1 - (1-p)^d] = \mathcal{O}(1)$ , we have  $\mathbb{E}[1/p^e(c_{11})] = \Theta(1/p)$  and  $\mathbb{E}[1/p^e(c_{01})] = \mathcal{O}(1)$ . Therefore, Condition 3.1 is satisfied.

Next, we show Condition 3.2 is also satisfied. Define a pairwise dependency indicator  $g_{ij}$  such that if  $g_{ij} = 0$ , then  $f(\mathbf{Z}, \mathbf{A}_i) \perp f(\mathbf{Z}, \mathbf{A}_j)$ , else let  $g_{ij} = 1$ . Note that  $g_{ij} = 1$  if  $C_{ij} \geq 1$  or  $A_{i,j} = 1$ . Then, we have  $\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} g_{ij} = o(N_v^2)$  because  $\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} I_{\{C_{ij}=0\}} \sim N_v^2$  and  $\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} I_{\{A_{i,j}=0\}} \sim N_v^2$ . Notice that

$$\begin{aligned}\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} |p_{ij}^e(c_k)/(p_i^e(c_k)p_j^e(c_k)) - 1| &= \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} g_{ij} |p_{ij}^e(c_k)/(p_i^e(c_k)p_j^e(c_k)) - 1| \\ &\leq \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} g_{ij} p_{ij}^e(c_k)/(p_i^e(c_k)p_j^e(c_k)) + \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} g_{ij}.\end{aligned}$$

It suffices to show

$$\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \frac{g_{ij} p_{ij}^e(c_k)}{p_i^e(c_k) p_j^e(c_k)} = o(N_v^2). \quad (\text{B.1})$$

By Young's inequality, we have

$$\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \frac{g_{ij} p_{ij}^e(c_k)}{p_i^e(c_k) p_j^e(c_k)} \leq \left( \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} g_{ij} \right)^{1/2} \left( \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \left[ \frac{p_{ij}^e(c_k)}{p_i^e(c_k) p_j^e(c_k)} \right]^2 \right)^{1/2}.$$

For the level  $c_{00}$ , we have

$$\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \left[ \frac{p_{ij}^e(c_{00})}{p_i^e(c_{00}) p_j^e(c_{00})} \right]^2 \leq \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \left[ \frac{1}{p_i^e(c_{00}) p_j^e(c_{00})} \right]^2 \leq \left( \sum_{i=1}^{N_v} \left[ \frac{1}{p_i^e(c_{00})} \right]^2 \right)^2$$

and  $\mathbb{E}[1/(p^e(c_{00}))^2] = \mathcal{O}(1)$ . Thus, we obtain

$$\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} |p_{ij}^e(c_{00}) / (p_i^e(c_{00}) p_j^e(c_{00})) - 1| = o(N_v^2).$$

Similarly, we can show (B.1) for other exposure levels. Therefore, Condition 3.1 is satisfied.

### Proof of Proposition 3.3

First, we compute the normalization parameter of the Pareto distribution with an exponential cutoff. By definitions, we have

$$C(\zeta, d_L, \lambda) \int_{d_L}^{N_v-1} e^{-\lambda x} x^{-(\zeta+1)} dx = 1,$$

$$C(\zeta, d_L, \lambda) \int_{d_L}^{N_v-1} x \cdot e^{-\lambda x} x^{-(\zeta+1)} dx = \bar{d}.$$

Note that, by integration by parts, we have

$$\begin{aligned} \int_{d_L}^{N_v-1} x \cdot e^{-\lambda x} x^{-(\zeta+1)} dx &= \frac{1}{\lambda} \cdot (d_L^{-\zeta} \cdot e^{-\lambda d_L} - (N_v - 1)^{-\zeta} \cdot e^{-\lambda(N_v-1)}) \\ &\quad - \frac{\zeta}{\lambda} \cdot \int_{d_L}^{N_v-1} e^{-\lambda x} x^{-(\zeta+1)} dx. \end{aligned}$$

Therefore, we obtain

$$C(\zeta, d_L, \lambda) = (\lambda \bar{d} + \zeta) / (d_L^{-\zeta} \cdot e^{-\lambda d_L} - (N_v - 1)^{-\zeta} \cdot e^{-\lambda(N_v-1)}), \quad (\text{B.2})$$

$$\int_{d_L}^{N_v-1} e^{-\lambda x} x^{-(\zeta+1)} dx = (d_L^{-\zeta} \cdot e^{-\lambda d_L} - (N_v - 1)^{-\zeta} \cdot e^{-\lambda(N_v-1)}) / (\lambda \bar{d} + \zeta). \quad (\text{B.3})$$

Next, we show  $d_L = \Theta(\bar{d})$ . By definitions of  $d_L$  and  $\bar{d}$ , we have  $d_L = \mathcal{O}(\bar{d})$ . Therefore, it suffices to show that  $d_L = \Omega(\bar{d})$ . We prove it by contradiction. Assume  $d_L = o(\bar{d})$ , then by (B.3), we have

$$\int_{d_L}^{N_v-1} e^{-\lambda x} x^{-(\zeta+1)} dx \sim d_L^{-\zeta} / (\lambda \bar{d} + \zeta). \quad (\text{B.4})$$

On the other hand, note that

$$\int_{d_L}^{N_v-1} e^{-\lambda x} x^{-(\zeta+1)} dx = \lambda^\zeta \left\{ \Gamma(-\zeta, \lambda d_L) - \Gamma(-\zeta, \lambda(N_v - 1)) \right\} \quad (\text{B.5})$$

where  $\Gamma(\cdot, \cdot)$  is the upper incomplete gamma function. Recall the properties of the upper incomplete gamma function,  $\Gamma(s, x) \rightarrow -x^s/s$  as  $x \rightarrow 0$  and  $s < 0$ , and  $\Gamma(s, x) \rightarrow x^{s-1}e^{-x}$  as  $x \rightarrow \infty$  (Jameson (2016), Olver (1997), Temme (2011)). Then

by (B.5), we obtain

$$\int_{d_L}^{N_v-1} e^{-\lambda x} x^{-(\zeta+1)} dx \sim d_L^{-\zeta}/\zeta, \quad (\text{B.6})$$

which is in contradiction to (B.4). Thus, we have  $d_L = \Theta(\bar{d})$ .

Then, we prove Proposition 3.3. By Proposition 3.1, it suffices to show Condition 3.1 and 3.2 are satisfied. Under Assumption 3.4 and  $p < \lambda$ , we obtain

$$\begin{aligned} \mathbb{E}[1/p^e(c_{10})] &= \frac{1}{p} \cdot C(\zeta, d_L, \lambda) \cdot \int_{d_L}^{N_v-1} (1-p)^{-x} e^{-\lambda x} x^{-(\zeta+1)} dx \\ &= \mathcal{O}\left(\frac{1}{p} \cdot C(\zeta, d_L, \lambda) \cdot d_L^{-(\zeta+1)} \cdot \int_{d_L}^{N_v-1} (1-p)^{-x} e^{-\lambda x} dx\right) \\ &= \mathcal{O}\left(\frac{1}{pd_L[\lambda + \log(1-p)]} \left\{ \left[\frac{1}{e^{\lambda(1-p)}}\right]^{d_L} - \left[\frac{1}{e^{\lambda(1-p)}}\right]^{N_v} \right\}\right) \\ &= \mathcal{O}(1/p). \end{aligned}$$

Similarly, we have  $\mathbb{E}[1/p^e(c_{00})] = \mathcal{O}(1)$ .

Next, we show  $\mathbb{E}[1/p^e(c_{11})] = \mathcal{O}(1/p)$  and  $\mathbb{E}[1/p^e(c_{01})] = \mathcal{O}(1)$ . Note that

$$\begin{aligned} \mathbb{E}[1/p^e(c_{11})] &= \frac{1}{p} \cdot C(\zeta, d_L, \lambda) \cdot \int_{d_L}^{N_v-1} \frac{1}{1 - (1-p)^x} e^{-\lambda x} x^{-(\zeta+1)} dx \\ &= \mathcal{O}\left(\frac{1}{p} \cdot C(\zeta, d_L, \lambda) \cdot d_L^{-(\zeta+1)} \cdot \int_{d_L}^{N_v-1} \frac{1}{1 - (1-p)^x} e^{-\lambda x} dx\right) \\ &= \mathcal{O}\left(\frac{1}{p} \cdot C(\zeta, d_L, \lambda) \cdot d_L^{-(\zeta+1)} \cdot \int_{d_L}^{N_v-1} e^{-\lambda x} dx\right) \\ &= \mathcal{O}\left(\frac{1}{p} \cdot C(\zeta, d_L, \lambda) \cdot d_L^{-(\zeta+2)}\right) \end{aligned}$$

$$= \mathcal{O}(1/p).$$

Similarly, we have  $\mathbb{E}[1/p^e(c_{01})] = \mathcal{O}(1)$ .

Analogous to the proof of Proposition 3.2, we can show Condition 3.2 is also satisfied.

## B.2 Proofs of propositions for noisy networks

### Proof of Proposition 3.4

In the observed network, the four exposure probabilities become:

$$\begin{aligned}\tilde{p}_i^e(c_{11}) &= p\{1 - (1 - p)^{\tilde{d}_i}\}, \\ \tilde{p}_i^e(c_{10}) &= p(1 - p)^{\tilde{d}_i}, \\ \tilde{p}_i^e(c_{01}) &= (1 - p)\{1 - (1 - p)^{\tilde{d}_i}\}, \\ \tilde{p}_i^e(c_{00}) &= (1 - p)^{\tilde{d}_i+1},\end{aligned}$$

where  $\tilde{d}_i$  is the degree of vertex  $i$  in the observed network.

(i) For all values  $i$  and  $c_k$ ,  $\mathbb{P}(\tilde{p}_i^e(c_k) > 0) \rightarrow 1$  as  $N_v \rightarrow \infty$ .

Note that  $\mathbb{P}(\tilde{p}_i^e(c_{10}) > 0) = \mathbb{P}(\tilde{p}_i^e(c_{00}) > 0) = 1$  and  $\mathbb{P}(\tilde{p}_i^e(c_{11}) > 0) = \mathbb{P}(\tilde{p}_i^e(c_{01}) > 0) = \mathbb{P}(\tilde{d}_i > 0)$ . Under Assumption 3.1 - 3.5, we have

$$\lim_{N_v \rightarrow \infty} \mathbb{P}(\tilde{d}_i > 0) = \lim_{N_v \rightarrow \infty} 1 - (1 - \alpha)^{N_v - 1 - d_i} \beta^{d_i} = 1.$$

Thus, for all values  $i$  and  $c_k$ , we have  $\mathbb{P}(\tilde{p}_i^e(c_k) > 0) \rightarrow 1$  as  $N_v \rightarrow \infty$ .

(ii) For all  $c_k$ ,  $\mathbb{E}[\sum_{i=1}^{N_v} I_{\{\tilde{p}_i^e(c_k) > 0\}} / \tilde{p}_i^e(c_k)] = o(N_v^2)$ .

Define  $\check{d}_i = \sum_{j=1}^{N_v} \tilde{A}_{j,i} A_{j,i}$ . We note that  $\tilde{d}_i = (\tilde{d}_i - \check{d}_i) + \check{d}_i$ , where  $\check{d}_i$  and  $\tilde{d}_i - \check{d}_i$  are two independent binomial random variables. Thus, we have  $\tilde{d}_i \sim \text{Binomial}(N_v -$

$1 - d_i, \alpha) + \text{Binomial}(d_i, 1 - \beta)$ . Then, we obtain

$$\begin{aligned}\mathbb{E}[(1 - p)^{\tilde{d}_i}] &= (1 - \alpha p)^{N_v - 1 - d_i} [1 - (1 - \beta)p]^{d_i}, \\ \text{Var}[(1 - p)^{\tilde{d}_i}] &= [1 - \alpha p(2 - p)]^{N_v - 1 - d_i} [1 - (1 - \beta)p(2 - p)]^{d_i} \\ &\quad - (1 - \alpha p)^{2(N_v - 1 - d_i)} [1 - (1 - \beta)p]^{2d_i}, \\ \mathbb{E}[(1 - p)^{-\tilde{d}_i}] &= \left(1 + \frac{\alpha p}{1 - p}\right)^{N_v - 1 - d_i} \left(1 + \frac{(1 - \beta)p}{1 - p}\right)^{d_i}.\end{aligned}$$

Since  $\mathbb{E}[(1 - p)^{-\tilde{d}_i}] = \mathcal{O}((1 - p)^{-d_i})$ , we have  $\mathbb{E}[1/\tilde{p}_i^e(c_{10})] = \mathcal{O}(1/p_i^e(c_{10}))$  and  $\mathbb{E}[1/\tilde{p}_i^e(c_{00})] = \mathcal{O}(1/p_i^e(c_{00}))$ . Therefore, by Proposition 3.2, we obtain  $\mathbb{E}[\sum_{i=1}^{N_v} I_{\{\tilde{p}_i^e(c_{10}) > 0\}}/\tilde{p}_i^e(c_{10})] = o(N_v^2)$  and  $\mathbb{E}[\sum_{i=1}^{N_v} I_{\{\tilde{p}_i^e(c_{00}) > 0\}}/\tilde{p}_i^e(c_{00})] = o(N_v^2)$ .

Note that  $\text{Var}[1 - (1 - p)^{\tilde{d}_i}] = o(1)$  and  $I_{\{\tilde{d}_i > 0\}} \xrightarrow{P} 1$ . Similar to the proof of Condition 3.1 satisfied in Proposition 3.2, we have

$$\lim_{N_v \rightarrow \infty} \mathbb{E}\left[\frac{p}{1 - (1 - p)^{\tilde{d}_i}} I_{\{\tilde{d}_i > 0\}}\right] = \lim_{N_v \rightarrow \infty} \frac{p}{\mathbb{E}[1 - (1 - p)^{\tilde{d}_i}]}.$$

Since  $\mathbb{E}[1 - (1 - p)^{\tilde{d}_i}] = \Theta(1 - (1 - p)^{d_i})$ , we have  $\mathbb{E}[I_{\{\tilde{p}_i^e(c_{11}) > 0\}}/\tilde{p}_i^e(c_{11})] = \mathcal{O}(1/p_i^e(c_{11}))$  and  $\mathbb{E}[I_{\{\tilde{p}_i^e(c_{01}) > 0\}}/\tilde{p}_i^e(c_{01})] = \mathcal{O}(1/p_i^e(c_{01}))$ . By Proposition 3.2, we obtain  $\mathbb{E}[\sum_{i=1}^{N_v} I_{\{\tilde{p}_i^e(c_{11}) > 0\}}/\tilde{p}_i^e(c_{11})] = o(N_v^2)$  and  $\mathbb{E}[\sum_{i=1}^{N_v} I_{\{\tilde{p}_i^e(c_{10}) > 0\}}/\tilde{p}_i^e(c_{10})] = o(N_v^2)$ .

(iii) For all  $c_k$ ,  $\mathbb{E}[\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} I_{\{\tilde{p}_i^e(c_k) > 0\}} I_{\{\tilde{p}_j^e(c_k) > 0\}} |\tilde{p}_{ij}^e(c_k)| / (\tilde{p}_i^e(c_k) \tilde{p}_j^e(c_k)) - 1] = o(N_v^2)$ .

Define a pairwise dependency indicator  $\tilde{g}_{ij}$  in the observed network such that if  $\tilde{g}_{ij} = 0$ , then  $f(\mathbf{Z}, \tilde{\mathbf{A}}_i) \perp\!\!\!\perp f(\mathbf{Z}, \tilde{\mathbf{A}}_j)$ , else let  $\tilde{g}_{ij} = 1$ . Then, we have  $\tilde{p}_{ij}^e(c_k) = \tilde{p}_i^e(c_k) \tilde{p}_j^e(c_k)$  if  $\tilde{g}_{ij} = 0$ . Thus, we obtain

$$\begin{aligned}
& \mathbb{E}\left[\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \tilde{g}_{ij} I_{\{\tilde{p}_i^e(c_k) > 0\}} I_{\{\tilde{p}_j^e(c_k) > 0\}} |\tilde{p}_{ij}^e(c_k) / (\tilde{p}_i^e(c_k) \tilde{p}_j^e(c_k)) - 1|\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} I_{\{\tilde{p}_i^e(c_k) > 0\}} I_{\{\tilde{p}_j^e(c_k) > 0\}} |\tilde{p}_{ij}^e(c_k) / (\tilde{p}_i^e(c_k) \tilde{p}_j^e(c_k)) - 1|\right].
\end{aligned}$$

By the triangle inequality, we have

$$\begin{aligned}
& \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \tilde{g}_{ij} I_{\{\tilde{p}_i^e(c_k) > 0\}} I_{\{\tilde{p}_j^e(c_k) > 0\}} |\tilde{p}_{ij}^e(c_k) / (\tilde{p}_i^e(c_k) \tilde{p}_j^e(c_k)) - 1| \\
&\leq \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \tilde{g}_{ij} I_{\{\tilde{p}_i^e(c_k) > 0\}} I_{\{\tilde{p}_j^e(c_k) > 0\}} \tilde{p}_{ij}^e(c_k) / (\tilde{p}_i^e(c_k) \tilde{p}_j^e(c_k)) + \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \tilde{g}_{ij} I_{\{\tilde{p}_i^e(c_k) > 0\}} I_{\{\tilde{p}_j^e(c_k) > 0\}}.
\end{aligned}$$

Note that  $\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \tilde{g}_{ij} I_{\{\tilde{p}_i^e(c_k) > 0\}} I_{\{\tilde{p}_j^e(c_k) > 0\}} \leq \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \tilde{g}_{ij}$ . Thus, it suffices to show  $\mathbb{E}[\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \tilde{g}_{ij} I_{\{\tilde{p}_i^e(c_k) > 0\}} I_{\{\tilde{p}_j^e(c_k) > 0\}} \tilde{p}_{ij}^e(c_k) / (\tilde{p}_i^e(c_k) \tilde{p}_j^e(c_k))] = o(N_v^2)$  and  $\mathbb{E}[\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \tilde{g}_{ij}] = o(N_v^2)$ .

First, we prove  $\mathbb{E}[\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \tilde{g}_{ij}] = o(N_v^2)$ . Let  $\tilde{C}_{ij}$  denote the number of common neighbors between vertex  $i$  and  $j$  in the observed network. Note that  $\tilde{g}_{ij} = 1$  if  $\tilde{C}_{ij} \geq 1$  or  $\tilde{A}_{i,j} = 1$ . Thus, it suffices to show  $\mathbb{E}[\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} I_{\{\tilde{C}_{ij}=0\}}] \sim N_v^2$  and  $\mathbb{E}[\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} I_{\{\tilde{A}_{i,j}=0\}}] \sim N_v^2$ . A direct computation yields to

$$\mathbb{E}\left[\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} I_{\{\tilde{A}_{i,j}=0\}}\right] = (1 - \alpha) \sum_{i=1}^{N_v} (N_v - 1 - d_i) + \beta \sum_{i=1}^{N_v} d_i \sim N_v^2.$$

Note that  $\tilde{C}_{ij} = \sum_{k=1}^{N_v} \tilde{A}_{k,i} \tilde{A}_{k,j} \sim \text{Binomial}(C_{ij}, (1-\beta)^2) + \text{Binomial}(d_i + d_j - 2C_{ij}, \alpha(1-\beta)) + \text{Binomial}(N_v - d_i - d_j + C_{ij} - 2, \alpha^2)$ , and the three binomial random variables

are independent. Then, we have

$$\begin{aligned} & \mathbb{E} \left[ \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} I_{\{\tilde{C}_{ij}=0\}} \right] \\ &= \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} [1 - (1 - \beta)^2]^{C_{ij}} [1 - \alpha(1 - \beta)]^{d_i + d_j - 2C_{ij}} (1 - \alpha^2)^{N_v - d_i - d_j + C_{ij} - 2}. \end{aligned}$$

Since  $\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} I_{\{C_{ij}=0\}} \sim N_v^2$ , we obtain  $\mathbb{E}[\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} I_{\{\tilde{C}_{ij}=0\}}] \sim N_v^2$ .

Next, we show  $\mathbb{E}[\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \tilde{g}_{ij} I_{\{\tilde{p}_i^e(c_k) > 0\}} I_{\{\tilde{p}_j^e(c_k) > 0\}} \tilde{p}_{ij}^e(c_k) / (\tilde{p}_i^e(c_k) \tilde{p}_j^e(c_k))] = o(N_v^2)$ .

By Young's inequality and Hölder's inequality, we have

$$\begin{aligned} & \mathbb{E} \left[ \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \frac{\tilde{g}_{ij} I_{\{\tilde{p}_i^e(c_k) > 0\}} I_{\{\tilde{p}_j^e(c_k) > 0\}} \tilde{p}_{ij}^e(c_k)}{\tilde{p}_i^e(c_k) \tilde{p}_j^e(c_k)} \right] \\ & \leq \left( \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \mathbb{E}(\tilde{g}_{ij}) \right)^{1/2} \left( \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \mathbb{E} \left[ \frac{I_{\{\tilde{p}_i^e(c_k) > 0\}} I_{\{\tilde{p}_j^e(c_k) > 0\}} \tilde{p}_{ij}^e(c_k)}{\tilde{p}_i^e(c_k) \tilde{p}_j^e(c_k)} \right]^2 \right)^{1/2}. \end{aligned}$$

For the level  $c_{00}$ , we have

$$\sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \mathbb{E} \left[ \frac{I_{\{\tilde{p}_i^e(c_{00}) > 0\}} I_{\{\tilde{p}_j^e(c_{00}) > 0\}} \tilde{p}_{ij}^e(c_{00})}{\tilde{p}_i^e(c_{00}) \tilde{p}_j^e(c_{00})} \right]^2 \leq \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} \mathbb{E} \left[ \frac{1}{\tilde{p}_i^e(c_{00}) \tilde{p}_j^e(c_{00})} \right]^2$$

and  $\mathbb{E}[(1/\tilde{p}_i^e(c_{00}) \tilde{p}_j^e(c_{00}))^2] = \mathcal{O}((1/p_i^e(c_{00}) p_j^e(c_{00}))^2)$ . Thus, we obtain

$$\mathbb{E} \left[ \sum_{i=1}^{N_v} \sum_{j \neq i}^{N_v} I_{\{\tilde{p}_i^e(c_{00}) > 0\}} I_{\{\tilde{p}_j^e(c_{00}) > 0\}} \tilde{p}_{ij}^e(c_{00}) / (\tilde{p}_i^e(c_{00}) \tilde{p}_j^e(c_{00})) \right] = o(N_v^2). \quad (\text{B.7})$$

Similarly, we can show (B.7) for other exposure levels. These complete the proof.

### Proof of Proposition 3.5

The proof of Proposition 3.5 is the same as that of Proposition 3.4.

### B.3 Proofs of theorems for noisy networks

First, we show the following lemmas that will be used in the proofs of theorems for noisy networks.

**Lemma B.1** *Let  $X_1, X_2, \dots$  and  $Y_1, Y_2, \dots$  be two nonnegative random variable sequences. Assume  $X_n I_{\{Y_n > 0\}}/Y_n$  is bounded and  $\mathbb{E}(Y_n) \neq 0$  for all  $n$ ,  $\text{Var}(X_n) = o(1)$ ,  $\text{Var}(Y_n) = o(1)$ , and  $I_{\{Y_n > 0\}} \xrightarrow{P} 1$  as  $n \rightarrow \infty$ . Then, we have  $\mathbb{E}[X_n I_{\{Y_n > 0\}}/Y_n] = \mathbb{E}(X_n)/\mathbb{E}(Y_n) + o(1)$ .*

**Proof.** Since  $\text{Var}(X_n) = o(1)$ , we have  $X_n \xrightarrow{L_2} \lim_{n \rightarrow \infty} \mathbb{E}(X_n)$ . This implies  $X_n \xrightarrow{P} \lim_{n \rightarrow \infty} \mathbb{E}(X_n)$ . Similarly, by continuous mapping theorem, we have  $I_{\{Y_n > 0\}}/Y_n \xrightarrow{P} \lim_{n \rightarrow \infty} 1/\mathbb{E}(Y_n)$ . Thus, we obtain  $X_n I_{\{Y_n > 0\}}/Y_n \xrightarrow{P} \lim_{n \rightarrow \infty} \mathbb{E}(X_n)/\mathbb{E}(Y_n)$ . Since  $X_n I_{\{Y_n > 0\}}/Y_n$  is bounded, we show  $\lim_{n \rightarrow \infty} \mathbb{E}[X_n I_{\{Y_n > 0\}}/Y_n] = \lim_{n \rightarrow \infty} \mathbb{E}(X_n)/\mathbb{E}(Y_n)$ . This completes the proof.  $\square$

**Lemma B.2** *Let  $T_1, T_2, \dots$ ,  $X_1, X_2, \dots$ , and  $Y_1, Y_2, \dots$  be three random variable sequences. Assume  $X_n = a + \mathcal{O}_p(g(n))$ ,  $Y_n = b + \mathcal{O}_p(h(n))$  as  $n \rightarrow \infty$ , where  $g(n) = o(a)$  and  $h(n) = o(b)$ . Let  $f_1(x, y, t), f_2(x, y, t), \dots$  be a sequence of function. And  $f_n(a, b, T_n) \xrightarrow{P} c$  as  $n \rightarrow \infty$ . In addition, we assume  $f_n(X_n, Y_n, T_n) - f_n(a, b, T_n) = o(1)$  when  $|X_n - a| \leq c_1 g(n)$  and  $|Y_n - b| \leq c_2 h(n)$ , where  $\mathbb{P}(|X_n - a| > c_1 g(n)) = o(1)$  and  $\mathbb{P}(|Y_n - b| > c_2 h(n)) = o(1)$ ,  $c_1 = \omega(1)$ ,  $c_2 = \omega(1)$ . Then, we have  $f_n(X_n, Y_n, T_n) \xrightarrow{P} c$  as  $n \rightarrow \infty$ .*

**Proof.** By triangle inequality, we have

$$|f_n(X_n, Y_n, T_n) - c| \leq |f_n(X_n, Y_n, T_n) - f_n(a, b, T_n)| + |f_n(a, b, T_n) - c|.$$

Then, for any  $\varepsilon_n > 0$ , we have

$$\begin{aligned}
& \mathbb{P}(|f_n(X_n, Y_n, T_n) - c| > \varepsilon_n) \\
& \leq \mathbb{P}(|f_n(X_n, Y_n, T_n) - f_n(a, b, T_n)| + |f_n(a, b, T_n) - c| > \varepsilon_n) \\
& \leq \mathbb{P}(|f_n(X_n, Y_n, T_n) - f_n(a, b, T_n)| > \varepsilon_n/2) + \mathbb{P}(|f_n(a, b, T_n) - c| > \varepsilon_n/2)
\end{aligned}$$

where the last step follows by the pigeonhole principle and the sub-additivity of the probability measure. Note that  $\lim_{n \rightarrow \infty} \mathbb{P}(|f_n(X_n, Y_n, T_n) - f_n(a, b, T_n)| > \varepsilon_n/2) = 0$  and  $\lim_{n \rightarrow \infty} \mathbb{P}(|f_n(a, b, T_n) - c| > \varepsilon_n/2) = 0$ . Thus, we obtain

$$\lim_{n \rightarrow \infty} \mathbb{P}(|f_n(X_n, Y_n, T_n) - c| > \varepsilon) = 0.$$

□

**Lemma B.3** *Let  $T_1, T_2, \dots$ ,  $X_1, X_2, \dots$ , and  $Y_1, Y_2, \dots$  be three random variable sequences. Assume  $X_n = a + \mathcal{O}_p(g(n))$ ,  $Y_n = b + \mathcal{O}_p(h(n))$  as  $n \rightarrow \infty$ , where  $g(n) = o(a)$  and  $h(n) = o(b)$ . Let  $f_1(x, y, t), f_2(x, y, t), \dots$  be a sequence of bounded function. And  $f_n(X_n, Y_n, T_n) - f_n(a, b, T_n) = o(1)$  when  $|X_n - a| \leq c_1 g(n)$  and  $|Y_n - b| \leq c_2 h(n)$ , where  $\mathbb{P}(|X_n - a| > c_1 g(n)) = o(1)$  and  $\mathbb{P}(|Y_n - b| > c_2 h(n)) = o(1)$ ,  $c_1 = \omega(1)$ ,  $c_2 = \omega(1)$ . Then, we have  $\mathbb{E}f_n(X_n, Y_n, T_n) - \mathbb{E}f_n(a, b, T_n) \rightarrow 0$  as  $n \rightarrow \infty$ .*

**Proof.** By the definition of the expectation, we have

$$\begin{aligned}
& \mathbb{E}f_n(X_n, Y_n, T_n) - \mathbb{E}f_n(a, b, T_n) \\
& = \mathbb{E} \left[ (f_n(X_n, Y_n, T_n) - f_n(a, b, T_n)) \cdot I_{\{|X_n - a| > c_1 g(n) \text{ or } |Y_n - b| > c_2 h(n)\}} \right] \\
& + \mathbb{E} \left[ (f_n(X_n, Y_n, T_n) - f_n(a, b, T_n)) \cdot I_{\{|X_n - a| \leq c_1 g(n) \text{ and } |Y_n - b| \leq c_2 h(n)\}} \right]
\end{aligned}$$

Since  $f_n$  is bounded, we have

$$\mathbb{E}f_n(X_n, Y_n, T_n) - \mathbb{E}f_n(a, b, T_n) \rightarrow 0$$

as  $n \rightarrow \infty$ . □

### Proof of Theorem 3.1

In the noisy network, Aronow and Samii estimators for  $\bar{y}(c_k)$  become

$$\begin{aligned} \tilde{y}_{A\&S}(c_{11}) &= \frac{1}{N_v} \sum_{i=1}^{N_v} I_{\{\tilde{d}_i > 0\}} I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{11}\}} \frac{1}{p[1 - (1-p)^{\tilde{d}_i}]} \left[ I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{11}\}} y_i(c_{11}) \right. \\ &\quad \left. + I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{10}\}} y_i(c_{10}) \right], \\ \tilde{y}_{A\&S}(c_{10}) &= \frac{1}{N_v} \sum_{i=1}^{N_v} I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{10}\}} \frac{1}{p(1-p)^{\tilde{d}_i}} \left[ I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{11}\}} y_i(c_{11}) \right. \\ &\quad \left. + I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{10}\}} y_i(c_{10}) \right], \\ \tilde{y}_{A\&S}(c_{01}) &= \frac{1}{N_v} \sum_{i=1}^{N_v} I_{\{\tilde{d}_i > 0\}} I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{01}\}} \frac{1}{(1-p)[1 - (1-p)^{\tilde{d}_i}]} \left[ I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{01}\}} y_i(c_{01}) \right. \\ &\quad \left. + I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{00}\}} y_i(c_{00}) \right], \\ \tilde{y}_{A\&S}(c_{00}) &= \frac{1}{N_v} \sum_{i=1}^{N_v} I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{00}\}} \frac{1}{(1-p)^{\tilde{d}_i+1}} \left[ I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{01}\}} y_i(c_{01}) \right. \\ &\quad \left. + I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{00}\}} y_i(c_{00}) \right]. \end{aligned}$$

The conditional biases for these estimators are

$$\begin{aligned} \text{Bias} \left[ \tilde{y}_{A\&S}(c_{11}) | \tilde{\mathbf{A}} \right] &= -\frac{1}{N_v} \sum_{i=1}^{N_v} I_{\{\tilde{d}_i > 0\}} \frac{(1-p)^{d_i} [1 - (1-p)^{\tilde{d}_i - d_i}]}{1 - (1-p)^{\tilde{d}_i}} \tau_i(c_{11}, c_{10}) \\ &\quad - \frac{1}{N_v} \sum_{i=1}^{N_v} I_{\{\tilde{d}_i = 0\}} y_i(c_{11}), \\ \text{Bias} \left[ \tilde{y}_{A\&S}(c_{10}) | \tilde{\mathbf{A}} \right] &= \frac{1}{N_v} \sum_{i=1}^{N_v} [1 - (1-p)^{d_i - \tilde{d}_i}] \tau_i(c_{11}, c_{10}), \end{aligned}$$

$$\begin{aligned} \text{Bias} \left[ \tilde{y}_{A\&S}(c_{01}) | \tilde{\mathbf{A}} \right] &= -\frac{1}{N_v} \sum_{i=1}^{N_v} I_{\{\tilde{d}_i > 0\}} \frac{(1-p)^{d_i} [1 - (1-p)^{\tilde{d}_i - d_i}]}{1 - (1-p)^{\tilde{d}_i}} \tau_i(c_{01}, c_{00}) \\ &\quad - \frac{1}{N_v} \sum_{i=1}^{N_v} I_{\{\tilde{d}_i = 0\}} y_i(c_{01}), \\ \text{Bias} \left[ \tilde{y}_{A\&S}(c_{00}) | \tilde{\mathbf{A}} \right] &= \frac{1}{N_v} \sum_{i=1}^{N_v} [1 - (1-p)^{d_i - \tilde{d}_i}] \tau_i(c_{01}, c_{00}). \end{aligned}$$

Based on the noisy network model, we have

$$\begin{aligned} \mathbb{E}[(1-p)^{d_i - \tilde{d}_i}] &= (1 - \beta p)^{d_i}, \\ \text{Var}[(1-p)^{d_i - \tilde{d}_i}] &= [1 - \beta p(2-p)]^{d_i} - (1 - \beta p)^{2d_i}, \\ \mathbb{E}[(1-p)^{\tilde{d}_i - d_i}] &= (1 - \alpha p)^{N_v - 1 - d_i}, \\ \text{Var}[(1-p)^{\tilde{d}_i - d_i}] &= [1 - \alpha p(2-p)]^{N_v - 1 - d_i} - (1 - \alpha p)^{2(N_v - 1 - d_i)}, \\ \mathbb{E}[(1-p)^{\tilde{d}_i}] &= (1 - \alpha p)^{N_v - 1 - d_i} [1 - (1 - \beta)p]^{d_i}, \\ \text{Var}[(1-p)^{\tilde{d}_i}] &= [1 - \alpha p(2-p)]^{N_v - 1 - d_i} [1 - (1 - \beta)p(2-p)]^{d_i} \\ &\quad - (1 - \alpha p)^{2(N_v - 1 - d_i)} [1 - (1 - \beta)p]^{2d_i}. \end{aligned}$$

For exposure levels  $c_{10}$  and  $c_{00}$ , by taking the expectation with respect to  $\tilde{\mathbf{A}}$ , we obtain

$$\begin{aligned}\text{Bias}\left[\tilde{y}_{A\&S}(c_{10})\right] &= \frac{1}{N_v} \sum_{i=1}^{N_v} [1 - (1 - \beta p)^{d_i}] \tau_i(c_{11}, c_{10}), \\ \text{Bias}\left[\tilde{y}_{A\&S}(c_{00})\right] &= \frac{1}{N_v} \sum_{i=1}^{N_v} [1 - (1 - \beta p)^{d_i}] \tau_i(c_{01}, c_{00}).\end{aligned}$$

Recall that  $p = o(1)$ . Then, we have  $\text{Var}[(1 - p)^{d_i - \check{d}_i}] = o(1)$ ,  $\text{Var}[(1 - p)^{\check{d}_i}] = o(1)$  and  $I_{\{\check{d}_i > 0\}} \xrightarrow{P} 1$ . For exposure levels  $c_{11}$  and  $c_{01}$ , by Lemma B.1, we have

$$\begin{aligned}\text{Bias}\left[\tilde{y}_{A\&S}(c_{11})\right] &= -\frac{1}{N_v} \sum_{i=1}^{N_v} \frac{(1 - p)^{d_i} [1 - (1 - \alpha p)^{N_v - 1 - d_i}]}{1 - (1 - \alpha p)^{N_v - 1 - d_i} (1 - (1 - \beta p)p)^{d_i}} \tau_i(c_{11}, c_{10}) + o(1), \\ \text{Bias}\left[\tilde{y}_{A\&S}(c_{01})\right] &= -\frac{1}{N_v} \sum_{i=1}^{N_v} \frac{(1 - p)^{d_i} [1 - (1 - \alpha p)^{N_v - 1 - d_i}]}{1 - (1 - \alpha p)^{N_v - 1 - d_i} (1 - (1 - \beta p)p)^{d_i}} \tau_i(c_{01}, c_{00}) + o(1).\end{aligned}$$

### Proof of Theorem 3.2

By the law of total variance, we have  $\text{Var}[\tilde{y}_{A\&S}(c_k)] = \text{Var}[\mathbb{E}(\tilde{y}_{A\&S}(c_k)|\tilde{\mathbf{A}})] + \mathbb{E}[\text{Var}(\tilde{y}_{A\&S}(c_k)|\tilde{\mathbf{A}})]$ . Thus, it suffices to show  $\text{Var}[\mathbb{E}(\tilde{y}_{A\&S}(c_k)|\tilde{\mathbf{A}})] = o(1)$  and  $\mathbb{E}[\text{Var}(\tilde{y}_{A\&S}(c_k)|\tilde{\mathbf{A}})] = o(1)$ .

$$(i) \text{Var}[\mathbb{E}(\tilde{y}_{A\&S}(c_k)|\tilde{\mathbf{A}})] = o(1).$$

Note that  $\text{Var}[\mathbb{E}(\tilde{y}_{A\&S}(c_k)|\tilde{\mathbf{A}})] = \text{Var}[\text{Bias}(\tilde{y}_{A\&S}(c_k)|\tilde{\mathbf{A}})]$ . For the exposure level  $c_{00}$ , we have

$$\begin{aligned}\text{Var}[\mathbb{E}(\tilde{y}_{A\&S}(c_{00})|\tilde{\mathbf{A}})] &= \frac{1}{N_v^2} \sum_{i=1}^{N_v} \text{Var}[1 - (1 - p)^{d_i - \check{d}_i}] \tau_i^2(c_{01}, c_{00}) \\ &\quad + \frac{1}{N_v^2} \sum_{i=1}^{N_v} \sum_{j \neq i} \text{Cov}((1 - p)^{d_i - \check{d}_i}, (1 - p)^{d_j - \check{d}_j}) \\ &\quad \cdot \tau_i(c_{01}, c_{00}) \tau_j(c_{01}, c_{00}).\end{aligned}$$

Note that  $\text{Var}[1 - (1 - p)^{d_i - \check{d}_i}] \leq 1$  and  $\text{Cov}((1 - p)^{d_i - \check{d}_i}, (1 - p)^{d_j - \check{d}_j}) = 0$  if  $A_{i,j} = 0$ . Thus, we have  $\text{Var}[\mathbb{E}(\tilde{y}_{A\&S}(c_{00})|\tilde{\mathbf{A}})] = o(1)$ . Similarly, we can show

$\text{Var}[\mathbb{E}(\tilde{y}_{A\&S}(c_k)|\tilde{\mathbf{A}})] = o(1)$  for other exposure levels.

(ii)  $\mathbb{E}[\text{Var}(\tilde{y}_{A\&S}(c_k)|\tilde{\mathbf{A}})] = o(1)$ .

For the exposure level  $c_{00}$ , we have

$$\begin{aligned}
& \text{Var} \left[ \tilde{y}_{A\&S}(c_{00}) \middle| \tilde{\mathbf{A}} \right] \\
&= \frac{1}{N_v^2} \left\{ \sum_{i=1}^{N_v} \left\{ \sum_{k' \in \{00,01\}} \check{p}_i^e(c_{00}, c_{k'}) \left[ 1 - \check{p}_i^e(c_{00}, c_{k'}) \right] \left[ \frac{y_i(c_{k'})}{\check{p}_i^e(c_{00})} \right]^2 \right. \right. \\
&\quad \left. \left. \check{p}_i^e(c_{00}, c_{00}) \check{p}_i^e(c_{00}, c_{01'}) \frac{y_i(c_{00})}{\check{p}_i^e(c_{00})} \frac{y_i(c_{01})}{\check{p}_i^e(c_{00})} \right\} \right. \\
&\quad \left. + \sum_{i=1}^{N_v} \sum_{j \neq i} \sum_{k' \in \{00,01\}} \sum_{k'' \in \{00,01\}} \left[ \check{p}_{ij}^e(c_{00}, c_{k'}, c_{00}, c_{k''}) \right. \right. \\
&\quad \left. \left. - \check{p}_i^e(c_{00}, c_{k'}) \check{p}_j^e(c_{00}, c_{k''}) \right] \frac{y_i(c_{k'})}{\check{p}_i^e(c_{00})} \frac{y_j(c_{k''})}{\check{p}_j^e(c_{00})} \right\}, \tag{B.8}
\end{aligned}$$

where

$$\begin{aligned}
\check{p}_{ij}^e(c_{k_1}, c_{k_2}) &= \sum_{\mathbf{z}} p_{\mathbf{z}} I_{\{f(\mathbf{z}, \tilde{\mathbf{A}}_i) = c_{k_1}\}} I_{\{f(\mathbf{z}, \mathbf{A}_i) = c_{k_2}\}}, \\
\check{p}_{ij}^e(c_{k_1}, c_{k_2}, c_{k_3}, c_{k_4}) &= \sum_{\mathbf{z}} p_{\mathbf{z}} I_{\{f(\mathbf{z}, \tilde{\mathbf{X}}_i) = c_{k_1}\}} I_{\{f(\mathbf{z}, \mathbf{x}_i) = c_{k_2}\}} I_{\{f(\mathbf{z}, \tilde{\mathbf{X}}_j) = c_{k_3}\}} I_{\{f(\mathbf{z}, \mathbf{x}_j) = c_{k_4}\}}
\end{aligned}$$

for  $k_1, k_2, k_3, k_4 = 1, 2, \dots, K$ ,  $i, j = 1, 2, \dots, N_v$ .

Note that  $\sum_{k' \in \{00,01\}} \check{p}_i^e(c_{00}, c_{k'}) = \check{p}_i^e(c_{00})$  and  $\sum_{k' \in \{00,01\}} \sum_{k'' \in \{00,01\}} \check{p}_{ij}^e(c_{00}, c_{k'}, c_{00}, c_{k''}) = \check{p}_{ij}^e(c_{00})$ . Then, (B.8) leads to

$$\text{Var} \left[ \tilde{y}_{A\&S}(c_{00}) \middle| \tilde{\mathbf{A}} \right] \leq \frac{C_1}{N_v^2} \sum_{i=1}^{N_v} \frac{1}{\check{p}_i^e(c_{00})} + \frac{C_2}{N_v^2} \sum_{i=1}^{N_v} \sum_{j \neq i} \frac{\check{g}_{ij} \check{p}_{ij}^e(c_{00})}{\check{p}_i^e(c_{00}) \check{p}_j^e(c_{00})},$$

where  $C_1$  and  $C_2$  are positive constants. By Proposition 3.4 and (B.7), we obtain  $\mathbb{E}[\text{Var}(\tilde{y}_{A\&S}(c_{00})|\tilde{\mathbf{A}})] = o(1)$ . Similarly, we can show  $\mathbb{E}[\text{Var}(\tilde{y}_{A\&S}(c_k)|\tilde{\mathbf{A}})] = o(1)$  for other exposure levels.

### Proof of Theorem 3.3

The proof of Theorem 3.3 is same as that of Theorem 3.2.

### Proof of Theorem 3.4

(i) Unbiasedness.

Note that  $\mathbb{P}(c_1/p \leq \hat{d}_i \leq c_2/p) \rightarrow 1$ ,  $c_1, c_2$  are constants, and  $\mathbb{E}[\tilde{y}_{A\&S,i} | \tilde{\mathbf{A}}]$  is bounded. Thus, it suffices to show  $\mathbb{E}[\tilde{y}_{\text{MME},i}(c_k) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}}] = y_i(c_k) + o(1)$ .

By the definition of  $\tilde{\mathbf{y}}_i$  in (4.1), we have

$$\begin{aligned}\tilde{y}_i(c_{11}) &= I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{11}\}} \left[ I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{11}\}} y_i(c_{11}) + I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{10}\}} y_i(c_{10}) \right], \\ \tilde{y}_i(c_{10}) &= I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{10}\}} \left[ I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{11}\}} y_i(c_{11}) + I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{10}\}} y_i(c_{10}) \right], \\ \tilde{y}_i(c_{01}) &= I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{01}\}} \left[ I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{01}\}} y_i(c_{01}) + I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{00}\}} y_i(c_{00}) \right], \\ \tilde{y}_i(c_{00}) &= I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{00}\}} \left[ I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{01}\}} y_i(c_{01}) + I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{00}\}} y_i(c_{00}) \right].\end{aligned}$$

By the definition of  $\mathbf{P}(d_i, \alpha, \beta)$  in (4.2), we have  $\mathbf{P}(d_i, \alpha, \beta) = \text{diag}(\mathbf{S}(d_i, \alpha, \beta), \mathbf{Q}(d_i, \alpha, \beta))$ , where

$$\begin{aligned}\mathbf{S}(d_i, \alpha, \beta) &= \begin{bmatrix} \mathbb{E}[I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{11}\}} I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{11}\}}] & \mathbb{E}[I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{11}\}} I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{10}\}}] \\ \mathbb{E}[I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{10}\}} I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{11}\}}] & \mathbb{E}[I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{10}\}} I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{10}\}}] \end{bmatrix}, \\ \mathbf{Q}(d_i, \alpha, \beta) &= \begin{bmatrix} \mathbb{E}[I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{01}\}} I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{01}\}}] & \mathbb{E}[I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{01}\}} I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{00}\}}] \\ \mathbb{E}[I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{00}\}} I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{01}\}}] & \mathbb{E}[I_{\{f(\mathbf{Z}, \tilde{\mathbf{A}}_i) = c_{00}\}} I_{\{f(\mathbf{Z}, \mathbf{A}_i) = c_{00}\}}] \end{bmatrix},\end{aligned}$$

and

$$\begin{aligned}
S_{11}(d_i, \alpha, \beta) &= p \left\{ 1 - (1-p)^{d_i} - (1-\alpha p)^{N_v-1-d_i} \cdot [(1-(1-\beta)p)^{d_i} - (1-p)^{d_i}] \right\}, \\
S_{12}(d_i, \alpha, \beta) &= p(1-p)^{d_i} [1 - (1-\alpha p)^{N_v-1-d_i}], \\
S_{21}(d_i, \alpha, \beta) &= p(1-\alpha p)^{N_v-1-d_i} [(1-(1-\beta)p)^{d_i} - (1-p)^{d_i}], \\
S_{22}(d_i, \alpha, \beta) &= p(1-p)^{d_i} (1-\alpha p)^{N_v-1-d_i}, \\
\mathbf{S}(d_i, \alpha, \beta) &= \frac{1-p}{p} \cdot \mathbf{Q}(d_i, \alpha, \beta).
\end{aligned}$$

Then,  $\tilde{\mathbf{y}}_{\text{MME},i}$  in (4.3) can be written as

$$\begin{aligned}
\tilde{\mathbf{y}}_{\text{MME},i}(c_{11}) &= S_{11}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \tilde{\mathbf{y}}_i(c_{11}) + S_{12}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \tilde{\mathbf{y}}_i(c_{10}), \\
\tilde{\mathbf{y}}_{\text{MME},i}(c_{10}) &= S_{21}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \tilde{\mathbf{y}}_i(c_{11}) + S_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \tilde{\mathbf{y}}_i(c_{10}), \\
\tilde{\mathbf{y}}_{\text{MME},i}(c_{01}) &= Q_{11}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \tilde{\mathbf{y}}_i(c_{01}) + Q_{12}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \tilde{\mathbf{y}}_i(c_{00}), \\
\tilde{\mathbf{y}}_{\text{MME},i}(c_{00}) &= Q_{21}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \tilde{\mathbf{y}}_i(c_{01}) + Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \tilde{\mathbf{y}}_i(c_{00}),
\end{aligned}$$

where  $\hat{\alpha}, \hat{\beta}$  are method-of-moments estimators obtained by Algorithm 3.1 in Chapter

3,  $\hat{d}_i = \frac{\hat{d}_i - (N_v - 1)\hat{\alpha}}{1 - \hat{\alpha} - \hat{\beta}}$  and

$$\begin{aligned}
S_{11}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) &= \frac{1}{p[1 - [1 - (1 - \hat{\beta})p]^{\hat{d}_i}]}, \\
S_{12}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) &= -\frac{1 - (1 - \hat{\alpha}p)^{N_v-1-\hat{d}_i}}{p(1 - \hat{\alpha}p)^{N_v-1-\hat{d}_i} [1 - [1 - (1 - \hat{\beta})p]^{\hat{d}_i}]}, \\
S_{21}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) &= -\frac{[1 - (1 - \hat{\beta})p]^{\hat{d}_i} - (1-p)^{\hat{d}_i}}{p(1-p)^{\hat{d}_i} [1 - [1 - (1 - \hat{\beta})p]^{\hat{d}_i}]}, \\
S_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) &= \frac{1 - (1-p)^{\hat{d}_i} - (1 - \hat{\alpha}p)^{N_v-1-\hat{d}_i} [(1 - (1 - \hat{\beta})p)^{\hat{d}_i} - (1-p)^{\hat{d}_i}]}{p(1-p)^{\hat{d}_i} (1 - \hat{\alpha}p)^{N_v-1-\hat{d}_i} [1 - [1 - (1 - \hat{\beta})p]^{\hat{d}_i}]}, \\
\mathbf{Q}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) &= \frac{p}{1-p} \cdot \mathbf{S}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}).
\end{aligned}$$

Thus, conditional expectations of  $\tilde{y}_{\text{MME},i}(c_k)$  are

$$\begin{aligned}
\mathbb{E}[\tilde{y}_{\text{MME},i}(c_{11})|\tilde{\mathbf{A}}] &= \left[ S_{11}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{11}, c_{11}) + S_{12}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{10}, c_{11}) \right] y_i(c_{11}) \\
&\quad + \left[ S_{11}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{11}, c_{10}) + S_{12}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{10}, c_{10}) \right] y_i(c_{10}), \\
\mathbb{E}[\tilde{y}_{\text{MME},i}(c_{10})|\tilde{\mathbf{A}}] &= \left[ S_{21}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{11}, c_{11}) + S_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{10}, c_{11}) \right] y_i(c_{11}) \\
&\quad + \left[ S_{21}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{11}, c_{10}) + S_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{10}, c_{10}) \right] y_i(c_{10}), \\
\mathbb{E}[\tilde{y}_{\text{MME},i}(c_{01})|\tilde{\mathbf{A}}] &= \left[ Q_{11}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{01}, c_{01}) + Q_{12}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{00}, c_{01}) \right] y_i(c_{01}) \\
&\quad + \left[ Q_{11}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{01}, c_{00}) + Q_{12}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{00}, c_{00}) \right] y_i(c_{00}), \\
\mathbb{E}[\tilde{y}_{\text{MME},i}(c_{00})|\tilde{\mathbf{A}}] &= \left[ Q_{21}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{01}, c_{01}) + Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{00}, c_{01}) \right] y_i(c_{01}) \\
&\quad + \left[ Q_{21}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{01}, c_{00}) + Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{00}, c_{00}) \right] y_i(c_{00}),
\end{aligned} \tag{B.9}$$

where

$$\begin{aligned}
\check{p}_i^e(c_{11}, c_{11}) &= p[1 - (1-p)^{d_i} - (1-p)^{\bar{d}_i} + (1-p)^{d_i+\bar{d}_i-\bar{d}_i}], \\
\check{p}_i^e(c_{10}, c_{10}) &= p(1-p)^{d_i+\bar{d}_i-\bar{d}_i}, \\
\check{p}_i^e(c_{11}, c_{10}) &= p(1-p)^{d_i}[1 - (1-p)^{\bar{d}_i-\bar{d}_i}], \\
\check{p}_i^e(c_{10}, c_{11}) &= p(1-p)^{\bar{d}_i}[1 - (1-p)^{d_i-\bar{d}_i}], \\
\check{p}_i^e(c_{01}, c_{01}) &= \frac{1-p}{p} \check{p}_i^e(c_{11}, c_{11}), \\
\check{p}_i^e(c_{00}, c_{00}) &= \frac{1-p}{p} \check{p}_i^e(c_{10}, c_{10}), \\
\check{p}_i^e(c_{01}, c_{00}) &= \frac{1-p}{p} \check{p}_i^e(c_{11}, c_{10}), \\
\check{p}_i^e(c_{00}, c_{01}) &= \frac{1-p}{p} \check{p}_i^e(c_{10}, c_{11}).
\end{aligned}$$

**Remark B.1** Note that  $\hat{\alpha}$ ,  $\hat{\beta}$  are method-of-moments estimators based on three observed networks  $\tilde{\mathbf{A}}$ ,  $\tilde{\mathbf{A}}_*$ , and  $\tilde{\mathbf{A}}_{**}$ . Thus, the expectations in (B.9) are actually conditional on  $\tilde{\mathbf{A}}$ ,  $\tilde{\mathbf{A}}_*$ , and  $\tilde{\mathbf{A}}_{**}$ . For notational simplicity, we omit  $\tilde{\mathbf{A}}_*$ ,  $\tilde{\mathbf{A}}_{**}$ .

Next, we compute the expectation of  $\mathbb{E}[\tilde{y}_{\text{MME},i}(c_k) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} | \tilde{\mathbf{A}}]$ . For the exposure level  $c_{11}$ , notice that

$$\begin{aligned} & S_{11}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{11}, c_{11}) + S_{12}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{10}, c_{11}) \\ &= \frac{[1 - (1-p)^{\hat{d}_i}]}{1 - [1 - (1-\hat{\beta})p]^{\hat{d}_i}} - \frac{(1-p)^{\hat{d}_i} [1 - (1-p)^{\hat{d}_i - \check{d}_i}]}{(1-\hat{\alpha}p)^{N_v-1-\hat{d}_i} [1 - [1 - (1-\hat{\beta})p]^{\hat{d}_i}]}. \end{aligned}$$

Since  $\text{Var}[(1-(1-\beta)p)^{\check{d}_i}] = o(1)$ , we have  $[1-(1-\beta)p]^{\check{d}_i} \xrightarrow{P} \lim_{N_v \rightarrow \infty} \mathbb{E}[(1-(1-\beta)p)^{\check{d}_i}]$ . By continuous mapping theorem, we obtain  $1 - [1 - (1-\beta)p]^{(\check{d}_i - \alpha(N_v-1))/(1-\alpha-\beta)} \xrightarrow{P} \lim_{N_v \rightarrow \infty} 1 - [1 - (1-\beta)p]^{-\alpha(N_v-1)/(1-\alpha-\beta)} (\mathbb{E}[(1-(1-\beta)p)^{\check{d}_i}])^{1/(1-\alpha-\beta)}$ . Note that  $\hat{\alpha} \xrightarrow{P} \alpha$  and  $\hat{\beta} \xrightarrow{P} \beta$ . Then, by lemma B.2, we have  $1 - [1 - (1-\hat{\beta})p]^{\hat{d}_i} \xrightarrow{P} \lim_{N_v \rightarrow \infty} 1 - [1 - (1-\beta)p]^{-\alpha(N_v-1)/(1-\alpha-\beta)} (\mathbb{E}[(1-(1-\beta)p)^{\check{d}_i}])^{1/(1-\alpha-\beta)}$ . By continuous mapping theorem, we obtain  $1/[1 - [1 - (1-\hat{\beta})p]^{\hat{d}_i}] \xrightarrow{P} \lim_{N_v \rightarrow \infty} 1/\mathbb{E}[1 - [1 - (1-\beta)p]^{(\check{d}_i - \alpha(N_v-1))/(1-\alpha-\beta)}]$ . Therefore, we have

$$\begin{aligned} & \mathbb{E}\left[\frac{[1 - (1-p)^{\hat{d}_i}]}{1 - [1 - (1-\hat{\beta})p]^{\hat{d}_i}} \cdot I_{\{\hat{d}_i = \Theta(1/p)\}}\right] \\ &= \frac{[1 - (1-p)^{\hat{d}_i}]}{1 - [1 - (1-\beta)p]^{-\alpha(N_v-1)/(1-\alpha-\beta)} (\mathbb{E}[(1-(1-\beta)p)^{\check{d}_i}])^{1/(1-\alpha-\beta)}} + o(1). \end{aligned}$$

Similarly, we can show

$$\begin{aligned} & \mathbb{E}\left[\frac{(1-p)^{\hat{d}_i} [1 - (1-p)^{\hat{d}_i - \check{d}_i}]}{(1-\hat{\alpha}p)^{N_v-1-\hat{d}_i} [1 - [1 - (1-\hat{\beta})p]^{\hat{d}_i}]}\right] \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} \\ &= \frac{(1-\alpha p)^{N_v-1-\hat{d}_i} [(1-(1-\beta)p)^{\hat{d}_i} - (1-p)^{\hat{d}_i}]}{1 - [1 - (1-\beta)p]^{-\alpha(N_v-1)/(1-\alpha-\beta)} (\mathbb{E}[(1-(1-\beta)p)^{\check{d}_i}])^{1/(1-\alpha-\beta)}} \\ & \cdot \frac{(1-\alpha p)^{N_v-1-\hat{d}_i} [(1-(1-\beta)p)^{\hat{d}_i} - (1-p)^{\hat{d}_i}]}{(1-\alpha p)^{N_v-1+\alpha(N_v-1)/(1-\alpha-\beta)} (\mathbb{E}[(1-\alpha p)^{-\check{d}_i}])^{1/(1-\alpha-\beta)}} + o(1). \end{aligned}$$

Then, direct computations yield to

$$\mathbb{E}[\{S_{11}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{11}, c_{11}) + S_{12}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{10}, c_{11})\} \cdot I_{\{\hat{d}_i = \Theta(1/p)\}}] = 1 + o(1).$$

Similarly, we have

$$\mathbb{E}[\{S_{11}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{11}, c_{10}) + S_{12}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \cdot \check{p}_i^e(c_{10}, c_{10})\} \cdot I_{\{\hat{d}_i = \Theta(1/p)\}}] = o(1).$$

Thus, we obtain  $\mathbb{E}[\tilde{y}_{\text{MME},i}(c_{11}) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}}] = y_i(c_{11}) + o(1)$ . Analogously, we can show  $\mathbb{E}[\tilde{y}_{\text{MME},i}(c_k) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}}] = y_i(c_k) + o(1)$  for other exposure levels.

(ii) Consistency.

Since  $\tilde{y}_{\text{MME}}(c_k)$  is an asymptotically unbiased estimator of  $\bar{y}(c_k)$ , it suffices to show  $\text{Var}(\tilde{y}_{\text{MME}}(c_k)) = o(1)$ . Note that  $\hat{\alpha} \xrightarrow{P} \alpha$ ,  $\hat{\beta} \xrightarrow{P} \beta$ ,  $\text{Var}(\tilde{y}_{\text{MME}}(c_k) | \tilde{\mathbf{A}})$  and  $\mathbb{E}(\tilde{y}_{\text{MME}}(c_k) | \tilde{\mathbf{A}})$  are bounded, by Lemma B.3, we have  $\text{Var}(\tilde{y}_{\text{MME}}(c_k))$  (unknown error rates)  $- \text{Var}(\tilde{y}_{\text{MME}}(c_k))$  (known error rates)  $= o(1)$ . Next, we compute  $\text{Var}(\tilde{y}_{\text{MME}}(c_k))$  when  $\alpha$  and  $\beta$  are known, i.e.,  $\hat{d}_i = \frac{\tilde{d}_i - (N_v - 1)\alpha}{1 - \alpha - \beta}$ .

By Cauchy-Schwarz inequality and the inequality  $2uv \leq u^2 + v^2$ , we have

$$\begin{aligned} \text{Var}[\tilde{y}_{\text{MME}}(c_k)] \leq & 2 \left\{ \text{Var} \left[ \frac{1}{N_v} \sum_{i=1}^{N_v} \tilde{y}_{\text{MME},i}(c_k) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} \right] \right. \\ & \left. + \text{Var} \left[ \frac{1}{N_v} \sum_{i=1}^{N_v} \tilde{y}_{\text{A\&S},i}(c_k) \cdot I_{\{\hat{d}_i = o(1/p) \text{ or } \omega(1/p)\}} \right] \right\}. \end{aligned}$$

Following the proof of Theorem 2, we can show  $\text{Var} \left[ \frac{1}{N_v} \sum_{i=1}^{N_v} \tilde{y}_{\text{A\&S},i}(c_k) \cdot I_{\{\hat{d}_i = o(1/p) \text{ or } \omega(1/p)\}} \right] = o(1)$ . Next, we prove  $\text{Var} \left[ \frac{1}{N_v} \sum_{i=1}^{N_v} \tilde{y}_{\text{MME},i}(c_k) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} \right] = o(1)$ .

For the exposure level  $c_{00}$ , we have

$$\tilde{y}_{\text{MME},i}(c_{00}) = Q_{21}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \tilde{p}_i(c_{01}) \cdot \tilde{y}_{\text{A\&S},i}(c_{01}) + Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \tilde{p}_i(c_{00}) \cdot \tilde{y}_{\text{A\&S},i}(c_{00}),$$

where  $Q_{21}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta})\tilde{p}_i(c_{01})$  and  $Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta})\tilde{p}_i(c_{00})$  are bounded when  $\hat{d}_i = \Theta(1/p)$ . Again, by Cauchy-Schwarz inequality and the inequality  $2uv \leq u^2 + v^2$ , we obtain

$$\begin{aligned} & \text{Var} \left[ \frac{1}{N_v} \sum_{i=1}^{N_v} \tilde{y}_{\text{MME},i}(c_{00}) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} \right] \\ & \leq 2 \left\{ \text{Var} \left[ \frac{1}{N_v} \sum_{i=1}^{N_v} Q_{21}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta})\tilde{p}_i(c_{01})\tilde{y}_{\text{A\&S},i}(c_{01}) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} \right] \right. \\ & \quad \left. + \text{Var} \left[ \frac{1}{N_v} \sum_{i=1}^{N_v} Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta})\tilde{p}_i(c_{00})\tilde{y}_{\text{A\&S},i}(c_{00}) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} \right] \right\}. \end{aligned} \quad (\text{B.10})$$

Thus, it suffices to show the two variances in (B.10) go to zero as  $N_v \rightarrow \infty$ . Here, we show  $\text{Var} \left[ \frac{1}{N_v} \sum_{i=1}^{N_v} Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta})\tilde{p}_i(c_{00})\tilde{y}_{\text{A\&S},i}(c_{00}) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} \right] = o(1)$ . The other one can be proved similarly.

By the law of total variance, we have

$$\begin{aligned} & \text{Var} \left[ \frac{1}{N_v} \sum_{i=1}^{N_v} Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta})\tilde{p}_i(c_{00})\tilde{y}_{\text{A\&S},i}(c_{00}) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} \right] \\ & = \text{Var} \left[ \mathbb{E} \left( \frac{1}{N_v} \sum_{i=1}^{N_v} Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta})\tilde{p}_i(c_{00})\tilde{y}_{\text{A\&S},i}(c_{00}) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} \mid \tilde{\mathbf{A}} \right) \right] \\ & \quad + \mathbb{E} \left[ \text{Var} \left( \frac{1}{N_v} \sum_{i=1}^{N_v} Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta})\tilde{p}_i(c_{00})\tilde{y}_{\text{A\&S},i}(c_{00}) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} \mid \tilde{\mathbf{A}} \right) \right]. \end{aligned}$$

Since  $Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta})\tilde{p}_i(c_{00}) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} = O(1)$ , we obtain

$$\begin{aligned} & \text{Var} \left[ \mathbb{E} \left( \frac{1}{N_v} \sum_{i=1}^{N_v} Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta})\tilde{p}_i(c_{00})\tilde{y}_{\text{A\&S},i}(c_{00}) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} \mid \tilde{\mathbf{A}} \right) \right] \\ & = \mathcal{O} \left( \text{Var} \left[ \mathbb{E} \left( \frac{1}{N_v} \sum_{i=1}^{N_v} \tilde{y}_{\text{A\&S},i}(c_{00}) \mid \tilde{\mathbf{A}} \right) \right] \right). \end{aligned}$$

Notice that  $\text{Var} \left[ \mathbb{E} \left( \frac{1}{N_v} \sum_{i=1}^{N_v} \tilde{y}_{\text{A\&S},i}(c_{00}) \mid \tilde{\mathbf{A}} \right) \right] = o(1)$  (shown in the proof of Theorem

2), we obtain

$$\text{Var} \left[ \mathbb{E} \left( \frac{1}{N_v} \sum_{i=1}^{N_v} Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \tilde{p}_i(c_{00}) \tilde{y}_{A\&S,i}(c_{00}) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} \mid \tilde{\mathbf{A}} \right) \right] = o(1).$$

Following the proof of Theorem 3.2, we can show

$$\begin{aligned} & \text{Var} \left( \frac{1}{N_v} \sum_{i=1}^{N_v} Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \tilde{p}_i(c_{00}) \tilde{y}_{A\&S,i}(c_{00}) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} \mid \tilde{\mathbf{A}} \right) \\ & \leq \frac{C_1}{N_v^2} \sum_{i=1}^{N_v} \frac{1}{\tilde{p}_i^e(c_{00})} + \frac{C_2}{N_v^2} \sum_{i=1}^{N_v} \sum_{j \neq i} \frac{\tilde{g}_{ij} \tilde{p}_{ij}^e(c_{00})}{\tilde{p}_i^e(c_{00}) \tilde{p}_j^e(c_{00})}, \end{aligned}$$

where  $C_1$  and  $C_2$  are positive constants. By Proposition 3.4 and (B.7), we obtain

$$\mathbb{E} \left[ \text{Var} \left( \frac{1}{N_v} \sum_{i=1}^{N_v} Q_{22}^{-1}(\hat{d}_i, \hat{\alpha}, \hat{\beta}) \tilde{p}_i(c_{00}) \tilde{y}_{A\&S,i}(c_{00}) \cdot I_{\{\hat{d}_i = \Theta(1/p)\}} \mid \tilde{\mathbf{A}} \right) \right] = o(1).$$

These complete the proof.

### Proof of Theorem 3.5

The proof of Theorem 3.5 is same as that of Theorem 3.4.

## B.4 Proofs of theorems in the generalized four-level exposure model

### Proof of Theorem 5.1

For  $m_i = 1$ , direct computations lead to the results in Theorem 5.1. Here, we show the case when  $m_i \geq 2$ . Note that, for all  $1 \leq x \leq d_i$ , we have (Das (2016))

$$\left( \frac{d_i}{x} \right)^x \leq \binom{d_i}{x} \leq \frac{d_i^x}{x!}.$$

For the exposure level  $c_{00'}$ , we obtain

$$\begin{aligned} p_i^e(c_{00'}) &\geq \frac{1-p}{(m_i-1)^{m_i-1}} \sum_{x=1}^{m_i-1} (d_i p)^x (1-p)^{d_i-x} + (1-p)^{d_i+1} \\ &\geq C_1 (1-p)^{d_i+1} \frac{(d_i p)^{m_i} (1-p)^{-(m_i-1)} - (1-p)}{p(d_i+1) - 1} \end{aligned}$$

and

$$\begin{aligned} p_i^e(c_{00'}) &\leq (1-p) \sum_{x=1}^{m_i-1} (d_i p)^x (1-p)^{d_i-x} + (1-p)^{d_i+1} \\ &\leq C_2 (1-p)^{d_i+1} \frac{(d_i p)^{m_i} (1-p)^{-(m_i-1)} - (1-p)}{p(d_i+1) - 1}, \end{aligned}$$

where  $C_1$  and  $C_2$  are positive constants. Thus, if  $p = o(1)$ , we have

$$p_i^e(c_{00'}) = \begin{cases} \Theta\left(\frac{(d_i p)^{m_i-1}}{e^{d_i p}}\right), & d_i = \omega(1/p), \\ \Theta(1), & d_i = \Theta(1/p), \\ \Theta(1), & d_i = o(1/p). \end{cases}$$

For the exposure level  $c_{01'}$ , we have

$$p_i^e(c_{01'}) \geq (1-p) \sum_{x=1}^{m_i-1} p^x (1-p)^{d_i-x} \geq C_3 (1-p) \frac{p^{m_i} (1-p)^{d_i-(m_i-1)} - p^{d_i+1}}{1-2p}$$

and

$$p_i^e(c_{01'}) \leq \frac{1-p}{m_i!} \sum_{x=m_i}^{d_i} (d_i p)^x (1-p)^{d_i-x} \leq C_4 (1-p) \frac{(d_i p)^{d_i+1} - (d_i p)^{m_i} (1-p)^{d_i-(m_i-1)}}{p(d_i+1) - 1},$$

where  $C_3$  and  $C_4$  are positive constants. Together with  $p_i^e(c_{01'}) = 1 - p - p_i^e(c_{00'})$ , we

obtain

$$p_i^e(c_{01'}) = \begin{cases} \Theta(1), & d_i = \omega(1/p), \\ \mathcal{O}(1) \text{ and } \Omega(p^{m_i}), & d_i = \Theta(1/p), \\ \mathcal{O}\left((d_i p)^{m_i}\right) \text{ and } \Omega(p^{m_i}), & d_i = o(1/p). \end{cases}$$

Note that  $p_i^e(c_{11'}) = p/(1-p) \cdot p_i^e(c_{01'})$  and  $p_i^e(c_{10'}) = p/(1-p) \cdot p_i^e(c_{00'})$ . The results for exposure levels  $c_{11'}$  and  $c_{10'}$  follow.

## B.5 Proofs for Pareto degree distribution without a cutoff

In this section, we show that assume a four-level exposure model and Bernoulli random assignment of treatment with  $p$ . In the inhomogeneous graph where the asymptotic degree distribution is the Pareto distribution with shape  $\zeta > 1$ , lower bound  $d_L$ , upper bound  $N_v - 1$  and mean  $\bar{d}$ , under Assumption 3.4, Condition 3.1 doesn't hold for levels  $c_{10}$ .

As  $N_v \rightarrow \infty$ , we have

$$\begin{aligned} \mathbb{E}[1/p^e(c_{10})] &= \frac{1}{p} \cdot \frac{\zeta d_L^\zeta}{1 - \left(\frac{d_L}{N_v - 1}\right)^\zeta} \cdot \int_{d_L}^{N_v - 1} (1-p)^{-x} x^{-(\zeta+1)} dx \\ &= \Theta\left(N_v \cdot \frac{d_L^\zeta}{p N_v} \cdot \int_{d_L}^{N_v - 1} (1-p)^{-x} x^{-(\zeta+1)} dx\right). \end{aligned} \quad (\text{B.11})$$

Note that  $\lim_{x \rightarrow \infty} (1-p)^{-x} x^{-(\zeta+1)} = \lim_{x \rightarrow \infty} p^{\zeta+1} e^{px}$ . By (B.11), we have

$$\mathbb{E}[1/p^e(c_{10})] = \Omega(N_v \cdot (d_L p)^\zeta \cdot e^{p N_v}). \quad (\text{B.12})$$

Next, we compute the order of  $d_L$ . By the definition of expectation, we have

$$\int_{d_L}^{N_v-1} x \cdot \frac{\zeta d_L^\zeta}{1 - \left(\frac{d_L}{N_v-1}\right)^\zeta} x^{-(\zeta+1)} dx = \begin{cases} \zeta d_L^\zeta \cdot \frac{\log\left(\frac{N_v-1}{d_L}\right)}{1 - \left(\frac{d_L}{N_v-1}\right)^\zeta}, & \text{if } \zeta = 1 \\ \frac{\zeta d_L}{\zeta-1} \cdot \frac{1 - \left(\frac{d_L}{N_v-1}\right)^{\zeta-1}}{1 - \left(\frac{d_L}{N_v-1}\right)^\zeta}, & \text{otherwise,} \end{cases}$$

Therefore, as  $N_v \rightarrow \infty$ , we obtain  $d_L = \Theta(\bar{d})$  when  $\zeta > 1$ . By (B.12), we have  $\mathbb{E}[1/p^e(c_{10})] = \Omega(N_v)$ . Therefore, condition 3.1 doesn't hold.

## References

- Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267.
- Abbott, S., Hellewell, J., Thompson, R. N., Sherratt, K., Gibbs, H. P., Bosse, N. I., Munday, J. D., Meakin, S., Doughty, E. L., Chun, J. Y., et al. (2020). Estimating the time-varying reproduction number of sars-cov-2 using national and subnational case counts. *Wellcome Open Research*, 5(112):112.
- Adam, D. C., Wu, P., Wong, J. Y., Lau, E. H. Y., Tsang, T. K., Cauchemez, S., Leung, G. M., and Cowling, B. J. (2020). Clustering and superspreading potential of sars-cov-2 infections in hong kong. *Nature Medicine*, 26(11):1714–1719.
- Ahmed, N. K., Neville, J., and Kompella, R. (2014). Network sampling: From static to streaming graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(2):7.
- Alvarez, R. M. and Levin, I. (2014). Uncertain neighbors: Bayesian propensity score matching for causal inference. Technical report, Technical report, California Institute of Technology, University of Georgia.
- Amaral, L. A. N., Scala, A., Barthélemy, M., and Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the national academy of sciences*, 97(21):11149–11152.
- An, W. (2010). 4. bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*, 40(1):151–189.
- Anderson, R. M. and May, R. (1991). Infectious diseases of humans. 1991. *New York: Oxford Science Publication Google Scholar*.
- Andersson, H. and Britton, T. (2012). *Stochastic epidemic models and their statistical analysis*, volume 151. Springer Science & Business Media.
- Aronow, P. M. and Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947.

- Arroyo Marioli, F., Bullano, F., Kučinskas, S., and Rondón-Moreno, C. (2020). Tracking  $r$  of covid-19: A new real-time estimation using the kalman filter. *Available at SSRN 3581633*.
- Athey, S., Eckles, D., and Imbens, G. W. (2018). Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240.
- Balachandran, P., Kolaczyk, E. D., and Viles, W. D. (2017). On the propagation of low-rate measurement error to subgraph counts in large networks. *The Journal of Machine Learning Research*, 18(1):2025–2057.
- Becker, N. G. and Britton, T. (1999). Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):287–307.
- Bhattacharya, R., Malinsky, D., and Shpitser, I. (2019). Causal inference under interference and network uncertainty. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 2019. NIH Public Access.
- Bloem-Reddy, B. and Orbanz, P. (2018). Random-walk models of network formation and sequential monte carlo methods for graphs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):871–898.
- Buono, C., Alvarez-Zuzek, L. G., Macri, P. A., and Braunstein, L. A. (2014). Epidemics in partially overlapped multiplex networks. *PloS one*, 9(3):e92200.
- Cevik, M., Marcus, J., Buckee, C., and Smith, T. (2020). Sars-cov-2 transmission dynamics should inform policy. *Available at SSRN 3692807*.
- Chang, J., Kolaczyk, E. D., and Yao, Q. (2020). Estimation of subgraph densities in noisy networks. *Journal of the American Statistical Association*, pages 1–14.
- Chatterjee, S. et al. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214.
- Chong, K. C., Cheng, W., Zhao, S., Ling, F., Mohammad, K. N., Wang, M., Zee, B. C., Wei, L., Xiong, X., Liu, H., et al. (2020). Transmissibility of coronavirus disease 2019 in chinese cities with different dynamics of imported cases. *PeerJ*, 8:e10350.
- Chowell, G., Castillo-Chavez, C., Fenimore, P. W., Kribs-Zaleta, C. M., Arriola, L., and Hyman, J. M. (2004a). Model parameters and outbreak control for sars. *Emerging Infectious Diseases*, 10(7):1258.

- Chowell, G., Hengartner, N. W., Castillo-Chavez, C., Fenimore, P. W., and Hyman, J. M. (2004b). The basic reproductive number of ebola and the effects of public health measures: the cases of congo and uganda. *Journal of theoretical biology*, 229(1):119–126.
- Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.
- Colman, E., Holme, P., Sayama, H., and Gershenson, C. (2019). Efficient sentinel surveillance strategies for preventing epidemics on networks. *PLoS computational biology*, 15(11):e1007517.
- Cori, A., Ferguson, N. M., Fraser, C., and Cauchemez, S. (2013). A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *Am J Epi*, 178(9).
- Cori, A., Kamvar, Z., Stockwin, J., Jombart, T., Thompson, R., and Dahlgvist, E. (2020). EpiEstim.
- Cox, D. R. and Cox, D. R. (1958). *Planning of experiments*, volume 20. Wiley New York.
- Craft, M. E., Volz, E., Packer, C., and Meyers, L. A. (2009). Distinguishing epidemic waves from disease spillover in a wildlife population. *Proceedings of the Royal Society B: Biological Sciences*, 276(1663):1777–1785.
- Cui, N., Chen, Y., and Small, D. S. (2013). Modeling parasite infection dynamics when there is heterogeneity and imperfect detectability. *Biometrics*, 69(3):683–692.
- Das, S. (2016). A brief note on estimates of binomial coefficients.
- Davoudi, B., Miller, J. C., Meza, R., Meyers, L. A., Earn, D. J., and Pourbohloul, B. (2012). Early real-time estimation of the basic reproduction number of emerging infectious diseases. *Physical Review X*, 2(3):031005.
- de Valpine, P., Paciorek, C., Turek, D., Michaud, N., Anderson-Bergman, C., Obermeyer, F., Wehrhahn Cortes, C., Rodríguez, A., Temple Lang, D., and Paganin, S. (2020a). *NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling*. R package version 0.10.1.
- de Valpine, P., Paciorek, C., Turek, D., Michaud, N., Anderson-Bergman, C., Obermeyer, F., Wehrhahn Cortes, C., Rodríguez, A., Temple Lang, D., and Paganin, S. (2020b). *NIMBLE User Manual*. R package manual version 0.10.1.

- de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26:403–413.
- Deng, X., Gu, W., Federman, S., du Plessis, L., Pybus, O. G., Faria, N., Wang, C., Yu, G., Bushnell, B., Pan, C.-Y., et al. (2020). Genomic surveillance reveals multiple introductions of sars-cov-2 into northern california. *Science*.
- Diekmann, O. and Heesterbeek, J. A. P. (2000). *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, volume 5. John Wiley & Sons.
- Drewe, J. A., Weber, N., Carter, S. P., Bearhop, S., Harrison, X. A., Dall, S. R., McDonald, R. A., and Delahay, R. J. (2012). Performance of proximity loggers in recording intra- and inter-species interactions: a laboratory and field-based validation study. *PLoS One*, 7(6):e39068.
- Eubank, S., Guclu, H., Kumar, V. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., and Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180.
- Fraser, C. (2007). Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. *PlosOne*, 2(8).
- Girolami, M. and Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.
- Grizzle, J. E. (1965). The two-period change-over design and its use in clinical trials. *Biometrics*, pages 467–480.
- Haario, H., Saksman, E., Tamminen, J., et al. (2001). An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242.
- Herrera, J. L., Srinivasan, R., Brownstein, J. S., Galvani, A. P., and Meyers, L. A. (2016). Disease surveillance on complex social networks. *PLoS computational biology*, 12(7):e1004928.
- Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.
- Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842.

- Jameson, G. (2016). The incomplete gamma functions. *The Mathematical Gazette*, 100(548):298–306.
- Jiang, X., Gold, D., and Kolaczyk, E. D. (2011). Network-based auto-probit modeling for protein function prediction. *Biometrics*, 67(3):958–966.
- Jiang, X. and Kolaczyk, E. D. (2012). A latent eigenprobit model with link uncertainty for prediction of protein–protein interactions. *Statistics in Biosciences*, 4(1):84–104.
- Juneau, C.-E., Briand, A.-S., Pueyo, T., Collazzo, P., and Potvin, L. (2020). Effective contact tracing for covid-19: A systematic review. *medRxiv*.
- Kao, R. R., Danon, L., Green, D. M., and Kiss, I. Z. (2006). Demographic structure and pathogen dynamics on the network of livestock movements in great britain. *Proceedings of the Royal Society B: Biological Sciences*, 273(1597):1999–2007.
- Kempton, R. and Lockwood, G. (1984). Inter-plot competition in variety trials of field beans (*vicia faba* l.). *The Journal of Agricultural Science*, 103(2):293–302.
- Kerr, C. C., Stuart, R. M., Mistry, D., Abeysuriya, R. G., Hart, G., Rosenfeld, K., Selvaraj, P., Nunez, R. C., Hagedorn, B., George, L., et al. (2020). Covasim: an agent-based model of covid-19 dynamics and interventions. *medRxiv*.
- Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data*. Springer.
- Kretzschmar, M. E., Rozhnova, G., Bootsma, M. C., van Boven, M., van de Wijgert, J. H., and Bonten, M. J. (2020). Impact of delays on effectiveness of contact tracing strategies for covid-19: a modelling study. *The Lancet Public Health*, 5(8):e452–e459.
- Kucharski, A. J., Wenham, C., Brownlee, P., Racon, L., Widmer, N., Eames, K. T., and Conlan, A. J. (2018). Structure and consistency of self-reported social contact networks in british secondary schools. *PloS one*, 13(7):e0200090.
- Latouche, P. and Robin, S. (2016). Variational bayes model averaging for graphon functions and motif frequencies inference in w-graph models. *Statistics and Computing*, 26(6):1173–1185.
- Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., and Lessler, J. (2020). The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172(9):577–582.
- Le, C. M. and Li, T. (2020). Linear regression and its inference on noisy network-linked data. *arXiv preprint arXiv:2007.00803*.

- Leavitt, S. V., Lee, R. S., Sebastiani, P., Horsburgh, C. R., Jenkins, H. E., and White, L. F. (2020). Estimating the relative probability of direct transmission between infectious disease patients. *International journal of epidemiology*.
- Li, T. and White, L. F. (2020). Bayesian back-calculation and nowcasting for line list data during the covid-19 pandemic. *medRxiv*.
- Li, W., Sussman, D. L., and Kolaczyk, E. D. (2020a). Estimation of the epidemic branching factor in noisy contact networks. *arXiv preprint arXiv:2002.05763*.
- Li, Y., Campbell, H., Kulkarni, D., Harpur, A., Nundy, M., Wang, X., Nair, H., for COVID, U. N., et al. (2020b). The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number ( $r$ ) of sars-cov-2: a modelling study across 131 countries. *The Lancet Infectious Diseases*.
- Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S.-m., Yuan, B., Kinoshita, R., and Nishiura, H. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *Journal of clinical medicine*, 9(2):538.
- Liu, Q.-H., Ajelli, M., Aleta, A., Merler, S., Moreno, Y., and Vespignani, A. (2018). Measurability of the epidemic reproduction number in data-driven contact networks. *Proceedings of the National Academy of Sciences*, 115(50):12680–12685.
- Luo, L., Liu, D., Liao, X.-l., Wu, X.-b., Jing, Q.-l., Zheng, J.-z., Liu, F.-h., Yang, S.-g., Bi, B., Li, Z.-h., et al. (2020). Modes of contact and risk of transmission in covid-19 among close contacts. *medRxiv*.
- Ma, Y., Jenkins, H., Sebastiani, P., Ellner, J., Jones-López, E., Dietze, R., Horsburgh, C., and White, L. (2020). Using cure models to estimate the serial interval of tuberculosis with limited follow-up. *Am J Epidemiol*.
- Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23.
- Martin, J., Wilcox, L. C., Burstedde, C., and Ghattas, O. (2012). A stochastic newton mcmc method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487.
- McClintock, B. T., Nichols, J. D., Bailey, L. L., MacKenzie, D. I., Kendall, W. L., and Franklin, A. B. (2010). Seeking a second opinion: uncertainty in disease ecology. *Ecology letters*, 13(6):659–674.

- Meredith, L. W., Hamilton, W. L., Warne, B., Houldcroft, C. J., Hosmillo, M., Jahun, A. S., Curran, M. D., Parmar, S., Caller, L. G., Caddy, S. L., et al. (2020). Rapid implementation of sars-cov-2 sequencing to investigate cases of health-care associated covid-19: a prospective genomic surveillance study. *The Lancet infectious diseases*, 20(11):1263–1272.
- Miller, D. A., Talley, B. L., Lips, K. R., and Campbell Grant, E. H. (2012). Estimating patterns and drivers of infection prevalence and intensity when detection is imperfect and sampling error occurs. *Methods in Ecology and Evolution*, 3(5):850–859.
- Neal, R. M. (2003). Slice sampling. *Annals of statistics*, pages 705–741.
- Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes (masters thesis); justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. excerpts english translation (reprinted). *Stat Sci*, 5:463–472.
- Olver, F. (1997). Asymptotics and special functions, ak peters. *Wellesley USA*.
- Parag, K. V. (2020). Improved estimation of time-varying reproduction numbers at low case incidence and between epidemic waves. *medRxiv*.
- Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925.
- Patki, N., Wedge, R., and Veeramachaneni, K. (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410. IEEE.
- Peters, P. J., Pontones, P., Hoover, K. W., Patel, M. R., Galang, R. R., Shields, J., Blosser, S. J., Spiller, M. W., Combs, B., Switzer, W. M., et al. (2016). Hiv infection linked to injection use of oxymorphone in indiana, 2014–2015. *New England Journal of Medicine*, 375(3):229–239.
- Poon, A. F., Gustafson, R., Daly, P., Zerr, L., Demlow, S. E., Wong, J., Woods, C. K., Hogg, R. S., Kraiden, M., Moore, D., et al. (2016). Near real-time monitoring of hiv transmission hotspots from routine hiv genotyping: an implementation case study. *The lancet HIV*, 3(5):e231–e238.
- Priebe, C. E., Sussman, D. L., Tang, M., and Vogelstein, J. T. (2015). Statistical inference on errorfully observed graphs. *Journal of Computational and Graphical Statistics*, 24(4):930–953.
- Reich, N., Lessler, J., Cummings, D., and Brookmeyer, R. (2009). Estimating incubation period distributions with coarse data. *Stat Med*, 28(22).

- Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *Journal of applied probability*, 44(2):458–475.
- Rosenbaum, P. R. (1999). Reduced sensitivity to hidden bias at upper quantiles in observational studies with dilated treatment effects. *Biometrics*, 55(2):560–564.
- Ross, R. (1916). An application of the theory of probabilities to the study of a priori pathometry.–part i. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character*, 92(638):204–230.
- Rubin, D., Huang, J., Fisher, B. T., Gasparrini, A., Tam, V., Song, L., Wang, X., Kaufman, J., Fitzpatrick, K., Jain, A., et al. (2020). Association of social distancing, population density, and temperature with the instantaneous reproduction number of sars-cov-2 in counties across the united states. *JAMA network open*, 3(7):e2016099–e2016099.
- Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference*, 25(3):279–292.
- Sansone, M., Andersson, M., Gustavsson, L., Andersson, L.-M., Nordén, R., and Westin, J. (2020). Extensive hospital in-ward clustering revealed by molecular characterization of influenza a virus infection. *Clinical Infectious Diseases*.
- Schafer, J. L. and Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods*, 13(4):279.
- Seemann, T., Lane, C., Sherry, N., Duchene, S., da Silva, A. G., Caly, L., Sait, M., Ballard, S. A., Horan, K., Schultz, M. B., et al. (2020). Tracking the covid-19 pandemic in australia using genomics. *medRxiv*.
- Smieszek, T., Burri, E. U., Scherzinger, R., and Scholz, R. W. (2012). Collecting close-contact social mixing data with contact diaries: reporting errors and biases. *Epidemiology & infection*, 140(4):744–752.
- Sussman, D. L. and Airoldi, E. M. (2017). Elements of estimation theory for causal effects in the presence of network interference. *arXiv preprint arXiv:1702.03578*.
- Temme, N. M. (2011). *Special functions: An introduction to the classical functions of mathematical physics*. John Wiley & Sons.
- Thompson, R. N., Stockwin, J. E., van Gaalen, R. D., Polonsky, J. A., Kamvar, Z. N., Demarsh, P. A., Dahlgvist, E., Li, S., Miguel, E., Jombart, T., Lessler, J., Cauchemez, S., and Cori, A. (2019). Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics*.
- Toulis, P. and Kao, E. (2013). Estimation of causal peer influence effects. In *International conference on machine learning*, pages 1489–1497.

- Trapman, P., Ball, F., Dhersin, J.-S., Tran, V. C., Wallinga, J., and Britton, T. (2016). Inferring  $r_0$  in emerging epidemics—the effect of common population structure is small. *Journal of the Royal Society Interface*, 13(121):20160288.
- Ugander, J., Karrer, B., Backstrom, L., and Kleinberg, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337.
- Vanhems, P., Barrat, A., Cattuto, C., Pinton, J.-F., Khanafer, N., Régis, C., Kim, B.-a., Comte, B., and Voirin, N. (2013). Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PloS one*, 8(9):e73970.
- Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P., Fu, H., et al. (2020). Estimates of the severity of covid-19 disease. *MedRxiv*.
- Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y., et al. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in wuhan, china. *Jama*, 323(11):1061–1069.
- White, L. F., Wallinga, J., Finelli, L., Reed, C., Riley, S., Lipsitch, M., and Pagano, M. (2009). Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza a/h1n1 pandemic in the usa. *Influenza and other respiratory viruses*, 3(6):267–276.
- Whittle, P. (1955). The outcome of a stochastic epidemic—a note on bailey’s paper. *Biometrika*, 42(1-2):116–122.
- Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., Niemeyer, D., Jones, T.C., Vollmar, P., Rothe, C., et al. (2020). Virological assessment of hospitalized patients with covid-2019. *Nature*, 581(7809):465–469.
- You, C., Deng, Y., Hu, W., Sun, J., Lin, Q., Zhou, F., Pang, C. H., Zhang, Y., Chen, Z., and Zhou, X.-H. (2020). Estimation of the time-varying reproduction number of covid-19 outbreak in china. *International Journal of Hygiene and Environmental Health*, page 113555.
- Young, J.-G., Cantwell, G. T., and Newman, M. (2020). Robust bayesian inference of network structure from unreliable data. *arXiv preprint arXiv:2008.03334*.

# CURRICULUM VITAE

