Title
Problems in Epidemic Inference on Complex Networks

Author
Kazemitabar Amirkolaei, Seyed Jalil

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Problems in Epidemic Inference on Complex Networks

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Seyed Jalil Kazemitabar Amirkolaei

2020

ABSTRACT OF THE DISSERTATION

Problems in Epidemic Inference on Complex Networks

by

Seyed Jalil Kazemitabar Amirkolaei
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2020
Professor Arash A. Amini, Chair

In this PhD dissertation, we study epidemics on networks of contacts through the lens of statistical inference. The current work is an attempt to infer the propagation parameters following the outset of an epidemic spread. My contributions rely on the progress on mathematical modeling of infectious outbreak, information diffusion, and viral habit formation. These achievements paved the path to forecast and contain the spread of infectious diseases and to optimize viral marketing campaigns. What distinguishes this work is the forensics view that aims to infer the network or the propagation parameters from the final stage of an epidemic. We study here multiple problems of this kind including epidemic source identification and epidemic network reconstruction. Such problems are NP-hard by nature and previous contributions are ad-hoc and inconclusive for realistic networks, either in size or in structure. This work proposes new methods that estimate the parameters of interest in polynomial time with arbitrary accuracy. We provide theoretical error bound guarantees for some of the solutions. We accompany the results with comparative simulations on popular networks from social media, urban infrastructure, and disease pandemics.

The dissertation of Seyed Jalil Kazemitabar Amirkolaei is approved.

Mason A. Porter

Qing Zhou

Yingnian Wu

Arash A. Amini, Committee Chair

University of California, Los Angeles

2020

To my parents . . .

to whom I owe my nature and my nurture,

and

to Zeinab . . .

who endured my distance

during the unjust "Muslim Travel Ban".

TABLE OF CONTENTS

I Epidemic Source Identification **3**

1 Variational Inference **4**

1.1 Introduction 4

1.2 Source detection in SI epidemics 7

1.2.1 Time and rate invariant analysis 8

1.2.2 Statistical Inference 8

1.3 Exact likelihood computation 10

1.4 Approximations 11

1.5 Simulations 14

2 Monte Carlo Estimation **20**

2.1 Introduction 20

2.1.1 The CH-SI dynamic 22

2.1.2 Flexibility of the model 23

2.2 Epidemic Inference from Complete Snapshots 24

2.2.1 Evaluation 26

2.3 Monte Carlo Estimation 27

2.3.1 Direct Monte Carlo sampling 28

2.3.2 Soft-Margin Monte Carlo sampling 29

2.3.3 Importance sampling 29

2.4 Consistency 31

2.4.1 Regularity assumptions 32

2.4.2 A note about the unbiased importance sampler 33

2.5 Simulations . . . 33
2.5.1 Comparison to the unbiased estimator . . . 38

II Epidemic Network Reconstruction 41

3 Monte Carlo Estimation . . . 42
3.1 Introduction . . . 42
3.2 The Epidemic Model . . . 45
3.3 Epidemic Inference from Cascade of Snapshots . . . 48
3.4 Monte Carlo Estimation . . . 50
3.4.1 Importance sampling . . . 50
3.4.2 Approximate inference . . . 51
3.5 Simulations . . . 52

III Appendices 55

4 Variational Source Identification . . . 56
4.1 Multi-source Extension . . . 56
4.2 Proofs . . . 56
4.2.1 Proof of Proposition 1 . . . 56
4.2.2 Proof of Proposition 2 . . . 57

5 Monte Carlo Source Identification . . . 62
5.1 Proofs . . . 62
5.1.1 Proof of Lemma 1 . . . 62
5.1.2 Proof of Theorem 1 . . . 62

5.1.3 Proof of Theorem 2 64
5.1.4 Proof of Corollary 1 66
5.1.5 Auxiliary lemmas 67

References **69**

LIST OF FIGURES

1.1 Plots of the expected relative rank versus the infection size for low-transitivity networks. 16

1.2 Plots of the expected relative rank versus the infection size for high-transitivity networks. 17

1.3 Runtime in seconds. 18

2.1 Plots of the expected relative rank versus the infection size for various networks: (left) DC-SBM, (right) Internet AS (bottom) US West Power Grid. 34

2.2 Plots of the expected relative rank versus the infection size (continued): (left) Wiki vote (right) UC64. 34

2.3 (Left) Average runtimes in seconds for infection size 100. (Right) Convergence rates for the expected relative rank as a function of the Monte Carlo sample size. (Bottom) Sampling variation of IS for a single epidemics of size 30. 37

2.4 Evaluating the unbiased estimator. Plots of the expected relative rank versus the infection size for various networks: (top left) DC-SBM, (top right) Internet AS, (bottom left) US West Power Grid, (bottom right) Wiki vote (bottom center) UC64. 39

2.5 Sampling variation of IS and unbiased IS for a single epidemics of size 30. 40

3.1 EPANET [Ros00] water distribution network 53

3.2 Plots of the bias and standard error on Epanet WDS 54

LIST OF TABLES

1.1 Network statistics 15

2.1 Network statistics 35

VITA

2006	Silver Medal, International Mathematics Olympiad, Slovenia
2009	B.S. Mathematics, Sharif University of Technology
2011	Research Assistant, Schaeffer Center for Health Policy & Economics, USC
2011–2013	Teaching Assistant, Economics Department, UC Berkeley
2013	Research Assistant, Boalt School of Law, UC Berkeley
2013	M.A. Economics, University of California, Berkeley
2014–2017	Data Scientist, Scientific Revenue, San Mateo, CA
2017–2018	Teaching Assistant, Statistics Department, UCLA
2017–2018	Research Assistant, Statistics Department, UCLA
2018	M.S. Statistics, University of California, Los Angeles

PUBLICATIONS

Seyed Jalal Kazemitabar, Farnoush Banaei-Kashani, Seyed Jalil Kazemitabar, and Dennis McLeod. "Efficient batch processing of proximity queries by optimized probing." *In Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 84-93. 2013.

Jalil Kazemitabar, Arash Amini, Adam Bloniarz, and Ameet S. Talwalkar. "Variable importance using decision trees." *In Advances in neural information processing systems*, pp. 426-435. 2017.

Jalil Kazemitabar and Javad Kazemitabar. "Measuring the conformity of distributions to Benford's law." *Communications in Statistics-Theory and Methods* (2019): 1-7.

S. Jalil Kazemitabar, and Arash A. Amini. "Approximate Identification of the Optimal Epidemic Source in Complex Networks." *In International Conference on Network Science*, pp. 107-125. Springer, Cham, 2020.

Introduction

The emergence of interest in complex relational systems has led to dramatic changes in many fields of science. The availability of detailed contact network data provides unique opportunities to study social relations, urban mobilization, the evolution of species, institutional politics, epidemiology, etc., at a finer level of details [EK10, BB04, JS08]. Working with complex networks, however, raises certain challenges in methodology beyond those of classical statistics. Measuring node centrality in large networks, the identification of cohesive groups and patterns of information propagation are among these challenges. Advances in these areas of network analysis can contribute simultaneously to many scientific fields.

We are interested in studying the network of links as a medium for propagation of information, disease, social habits, pollution, etc. This topic which is often referred to as "network diffusion" or "network epidemic" modeling has applications in a number of real-world problems such as the study of viral marketing [KKT03], voter behavior [SR05], cascades of financial bankruptcies [RBP13], and the containment of contagious diseases [LDP15].

In this PhD thesis, we present novel methodological contributions to the study of epidemic inference problems. We cover a class of problems that aim to retrieve the hidden parameters of the propagation dynamics given static snapshots of the epidemics. This type of inference has been shown to have many real-world applications. Examples include inferring the strength of ties in a social blogging network from the shared stories, inferring the important variables that contribute to the propagation of Ebola in West Africa using the time-free data of the epidemics [DCB17], and identifying the corrupt ingredient in a food distribution system after noticing a list of contaminated items in retail stores [MKS14a]. In the course of this thesis, we propose a unifying formulation for this class of problems and introduce efficient Monte Carlo and variational inference solutions.

The algorithms developed here can be used to efficiently solve many types of epidemic inference problems including: 1) retrieving the parameters of the epidemic, such as the rate, the duration and the source of the spread, 2) assessing whether the spread is contagious or random, and 3) reconstructing the epidemic network.

Cautions should be made if one wants to apply our methods directly to the real-world problems in epidemiology or social sciences. The computational break-throughs introduced here build on modeling assumptions about the nature of the epidemic spread. Among the myriad models of viral behavior we picked a version of Susceptible-infected dynamics for our setup. To the extent practitioners believe this is a valid assumption to the application at hand, they can enjoy the accuracy and performance of our methods.

Acknowledgment

We would like to thank Mason A. Porter for providing the Facebook-100 dataset.

Part I

Epidemic Source Identification

CHAPTER 1

Variational Inference

1.1 Introduction

Modern transportation networks have had profound effects on geographical spread of infectious diseases [CH04, Coh00] giving rise to complicated epidemic evolutions [CBB06]. These evolutions can be modeled as dynamic processes on transportation networks. The epidemic spread on networks can take other forms, such as outbreaks of foodborne diseases [SAS98], intercontinental cascade of failures among financial institutions [EGJ14, AOT15], computer malware propagation on the internet and mobile networks [Kon08, FLJ07] spread of targeted fake news [SCV17, SCV18] and rumors [FAE14] on social media, especially during presidential elections [SJD17, JCG17, AG17]. In response to an adverse pandemics on a network, it is critical to trace back sources to enable appropriate prevention and containment of the spread [Org08]. Inferential methods have been developed to locate the source of foodborne diseases [MKS14b, HF19] and influenza pandemics [SCW16, PJZ19]. In the context of online social networks, the spread of misinformation can be limited by the identification of influential users [PMA14, KGH10]. Source recovery can also be used to assess the power of diffusions in generating anonymity in network protocols [BFV17].

The epidemic source identification problem has received considerable attention in the past decade. Given a snapshot of the infected nodes in a network, the task is to discover who has originated the epidemic. Since the seminal work of Shah and Zaman [SZ11a], numerous attempts have been made to address the question and its extensions [FC12, LMO14, ZY16a, LT12, NGM16a, PVF12]. By now, there are multiple methods that show satisfactory results in limited experimental setups or have proven guarantees in restricted

network topologies [JWY17]. However, identifying the source under general conditions still remains a difficult task. The problem of optimal recovery appears to be NP-hard in infection size [NGM16a, LTG10]. The theoretical guarantees for optimal and consistent recovery are restricted to regular infinite trees [SZ11a, ZY16a], and as we show in this paper, the popular and well-cited methods are quite unreliable in a wide range of real networks.

Source identification has remained largely unsolved and poorly understood for real complex networks [JWY17]. As we will show through experiments in Section 1.5, in real networks, even the optimal Bayes estimator applied to small infected sets has difficulty narrowing down to the true source. It is thus important to recover as much information from the likelihood of the model as possible. We develop techniques for computing the full likelihood of the infection, as opposed to identifying the most likely sample-path [ZY16a]. Moreover, we fully exploit the information from the boundary of the infection set, in addition to the structure inside the infected subgraph. We develop all these ideas without restricting the structure of the network to trees. Our framework also easily extends to the case where there are multiple infecting sources (Appendix 4.1).

In this paper, we develop statistical algorithms that outperform the state-of-the-art in a wide range of network topologies. Our contributions are distinct in several ways:

1. Our methods are parameter-free, meaning that they do not require knowing the duration of the epidemic or how fast it grows [LMO14, ALS15].
2. We show that the exact maximum likelihood estimator (MLE) of the source—or equivalently the Bayes optimal solution under uniform prior—can be written as a dynamic programming (DP), with easily computable coefficients based on the adjacency matrix of the network.
3. We develop two schemes to approximate the DP: an efficient greedy elimination (GE), and a novel mean-field approximation (MFA) of the likelihood, computed by solving a linear system. MFA and GE both perform well in naturally occurring networks, extend directly to heterogeneous infection probabilities, and are scalable, while competing methods fail to succeed in general topology.
4. Our approximations are more disciplined than existing approaches. They do not impose

restrictions on the topology of the network. Nor do they appeal to the partial likelihood of the candidate infecting sets. This is in contrast to the use of spanning trees to deal with general topologies [SZ11a, PLS18] or the *path-based approaches* that rely on the likelihood of individual paths from potential sources to the infected set [ZY16a].

We will show that when applied to real networks, both approximation schemes (MFA and GE) outperform various geometric and spectral approaches, most of which perform no better than random guessing. We also show that even for basic models of real networks, e.g., models with community structure, most existing methods dramatically fail. The improvement in performance is most significant for the networks with many cycles, including social networks that are known to have high transitivity. In terms of computational efficiency, both the greedy and mean-field approximations are superior to the state-of-the-art likelihood-based and spectral approaches and comparable to centrality-based methods. In addition, the mean-field algorithm is easily parallelizable through standard linear algebraic routines and can be used to tackle very large-scale epidemics on real networks.

Related work. Most of the existing literature on the source identification problem are based on an SIR dynamic where the infection spreads with an exponential rate proportional to the number of infected neighbors. All nodes are *susceptible* to the infection and once *infected* may *recover* with a fixed exponential rate [KMS17]. Moreover, the spread of infection through edges are mutually independent. Different variations of SIR may assume that no recovery is possible (SI) or the recovered is not immune to iterated infections (SIS).

Shah and Zaman [SZ11a] considered the SI dynamics and proposed the Rumor Centrality (RC), which counts the *permitted permutations*, a.k.a. infection paths, inside the infected subgraph. Their linear time algorithm is an optimal estimator in regular trees and enjoys strong theoretical properties in such idealized settings [KL17a]. Zhou and Ying [ZY16a] consider SIR dynamics on a tree and show that the most likely infection path is rooted at a Jordan center (JC) of the infected set O, that is, a node with minimal eccentricity (i.e., minimal maximum distance to other nodes). It has been shown [ZY16a, KL17a] that in regular trees, eccentricity ranking generates, with high probability, a confidence set containing

the true source, whose size does not grow with the infection size.

The Dynamic Message Passing (DMP) was proposed in [LMO14] as an approximation of the maximum likelihood estimator in discrete SIR epidemics, by approximating the probability of an infected set, as the product of the marginal probabilities of infection for each node (i.e., a form of pseudo-likelihood). Despite compelling performance, DMP is computationally intensive and impractical for large networks with moderately dense structures, even for small infection sets. A spectral algorithm, called Dynamical Age (DA) was introduced in [FC12], based on how sensitive the maximum eigenvalue of the Laplacian matrix is to the elimination of each node in the infection set. The algorithm was mainly developed to discover the initial node in a growing preferential attachment model. Another spectral method for the discrete SI model is proposed in [PVF12].

1.2 Source detection in SI epidemics

We consider a continuous-time heterogeneous susceptible-infected (SI) epidemic [KMS17] with rate of infection β, on a static weighted (directed) network $G(V, E)$ with known edge set E and $V = [n]$. At time zero, all nodes but the source are in the susceptible state. Infection is a terminal state and susceptible nodes are exposed to the infection at an exponential rate proportional to the number of their infected neighbors. More precisely, given that nodes I are infected at some time t, we run exponential clocks $T_j \sim \mathrm{Exp}(\beta \,\mathrm{vol}(I, j))$ for all $j \in I^c$ and the first to expire determines the next infected node: If $\boldsymbol{j}^* = \mathrm{argmin}_j T_j$, then the dynamics move to the infected set $I \cup \{\boldsymbol{j}^*\}$ at time $t + T_{\boldsymbol{j}^*}$. It is clear that the contagion will eventually spread through the entire graph.

The infection source or patient zero, denoted as $\boldsymbol{i}_*$, is unknown. What we observe is a snapshot of the contagion at some time t, meaning the entire set of infected nodes at that time, which we denote by O. The objective is to find $\boldsymbol{i}_* \in O$ or form a confidence set for $\boldsymbol{i}_*$ with desired false exclusion probability. Our focus here will be on the single source setting, but the analysis is extensible to the multi-source setting (cf. Section 4.1).

Notation. We write $A \in [0, 1]^{n \times n}$ for the weighted (asymmetric) adjacency matrix of the

network and $\mathrm{vol}(I, J) := \sum_{i \in I, j \in J} A_{ij}$ for the volume of a cut in the network between subsets $I, J \subset [n]$ of nodes. For singleton subsets, we often drop the braces, e.g., $\mathrm{vol}(I, j) := \mathrm{vol}(I, \{j\})$ and $O \setminus j = O \setminus \{j\}$.

1.2.1 Time and rate invariant analysis

We start by examining the probability of observing a particular set of infected nodes given a starting source. Let us introduce a parameter-free formulation of the problem (i.e. not dependent on rate β and time t) that will be the foundation for our analysis of the continuous SI dynamics.

Suppose that, at some point in time, the infection reaches $I \subset [n]$. Let $O \subset [n]$ be some superset of I. We are interested in computing $\rho_{I \to O}$, the chance that all the nodes in O are infected before any node outside. More precisely, let

$$\rho_{I \to O} := \mathbb{P}\Big(O \text{ is infected before } O^c \mid I \text{ is infected}\Big). \tag{1.1}$$

We refer to $\rho_{I \to O}$ as the *transition probabilities*. Note that these transition probabilities are independent of the infection source. Given that in a snapshot of the contagion, nodes I are infected, $\rho_{I \to O}$ determines how likely it is that in some future snapshot, O is the set of infected nodes. The Markov property of (continuous-time) SI dynamics allows us to define $\rho_{I \to O}$ without reference to the source, or the time of the first snapshot. We will also show that these probabilities do not require the knowledge of the infection rate or the time of the second snapshot.

1.2.2 Statistical Inference

Given the observed (random) infected set O, the function $I \mapsto \rho_{I \to O}$ is *the likelihood* of the model. Writing $L_O(I) := \rho_{I \to O}$ for this likelihood, we observe that $L_O(I) = 0$ for all I not contained in O. So, we can restrict $L(\cdot)$ to all subsets of O. When dealing with the single-source setup, we restrict the parameter space to $I = \{i\}$ and with some abuse of notation write $\rho_{i \to O}$ for $\rho_{\{i\} \to O}$, and $L_O(i) = \rho_{i \to O}, i \in [n]$ for the likelihood.

We can further consider a Bayesian setup by putting a uniform prior on the source (i.e., uniform over $[n]$). The Bayesian setup allows us to consider various notions of optimality by changing the loss function. Letting $\boldsymbol{i}_*$ be the random initial source, we have a joint distribution on $(\boldsymbol{i}_*, O)$. Then the posterior probability that the source is i, given that we observed infected nodes O is

$$p_i := \mathbb{P}(\boldsymbol{i}_* = i \mid O) = \frac{\rho_{i\to O}}{\sum_{j\in O} \rho_{j\to O}} 1\{i \in O\}.$$

Therefore, the maximum a posteriori (MAP) estimate of the source is $\boldsymbol{i}^*_{\mathrm{MAP}} = \mathrm{argmax}_i\ \rho_{i\to O}$ which minimizes the probability of error. That is, i^*_{MAP} minimizes $\mathbb{P}(\hat{\boldsymbol{i}} \neq \boldsymbol{i}_*)$ for any estimator $\hat{\boldsymbol{i}} = \hat{\boldsymbol{i}}(O)$. In some applications, the graph geodesic distance (d_G) to the source determines the error of estimation. In that case, the Bayes optimal estimator is $\boldsymbol{i}^*_{\mathrm{dist}} = \mathrm{argmin}_i \sum_{j\in O} \mathrm{dist}_G(i,j)\, \rho_{j\to O}$. It is not hard to see that $\boldsymbol{i}^*_{\mathrm{dist}}$ minimizes $\mathbb{E}[d_G(\hat{\boldsymbol{i}}, \boldsymbol{i}_*)]$ among all possible estimators $\hat{\boldsymbol{i}}$.

A third choice is to output a ranking instead of a single source. In this case, an estimator is formally a permutation $\hat{\boldsymbol{\sigma}} = \hat{\boldsymbol{\sigma}}_O$ on $[n]$, suppressing the dependence on O for simplicity. We can then consider the *rank loss* $\ell(\hat{\boldsymbol{\sigma}}, \boldsymbol{i}_*) = \hat{\boldsymbol{\sigma}}(\boldsymbol{i}_*)$, and we call the associated risk the *expected (source) rank* $= \mathbb{E}\hat{\boldsymbol{\sigma}}(\boldsymbol{i}_*)$. The corresponding optimal Bayes estimator is obtained by minimizing the posterior risk:

$$\hat{\boldsymbol{\sigma}}^* := \operatorname*{argmin}_{\sigma:[n]\to[n]}\ \mathbb{E}[\sigma(\boldsymbol{i}_*) \mid O].$$

Noting that $\mathbb{E}[\sigma(\boldsymbol{i}_*) \mid O] = \sum_i \sigma(i)\, p_i$, the optimal estimator in this case is the ranking that sorts p_i into descending order, i.e., $\hat{\boldsymbol{\sigma}}^*(j_i) = i$ where $p_{j_1} \geq p_{j_2} \geq \cdots \geq p_{j_n}$.

Remark 1. The distance loss might be suitable in some applications, but in general it is a poor measure if the goal is to reveal the actual source. This is especially true in small world networks, including most social networks, where the expected distance between any pair of nodes is small. On the other extreme, in terms of the precision in recovering the source, is the zero-one loss which is too stringent. The rank loss can be considered a more robust version of the zero-one loss, and it will be our main evaluation measure.

1.3 Exact likelihood computation

The Bayesian estimators introduced in Section 1.2.2 require us to evaluate the posterior probabilities (p_i), or equivalently the likelihood values $\rho_{j\to O}$ for all $j \in O$. The main difficulty of the source identification problem is that computing the likelihood is itself challenging. We now develop exact equations that allow us to recursively compute the likelihood values $L_O(I)$ for all subsets $I \subset O$.

Dynamic programming. To begin, note that $\rho_{O\to O} = 1$ for any $O \subset [n]$. In addition, $\rho_{I\to O} = 1$ whenever O corresponds to a connected component of G. We develop two dynamic programming expressions for $\rho_{I\to O}$ for general $I \subset O$:

Proposition 1. *For $I \subset O \subset [n]$, the probabilities $\rho_{I\to J}$ defined in* (1.1) *satisfy the forward program*

$$\rho_{I\to O} = \sum_{j\in O\setminus I} \frac{\mathrm{vol}(I,j)}{\mathrm{vol}(I,I^c)}\,\rho_{I\cup j\to O} \tag{1.2}$$

and the backward program

$$\rho_{I\to O} = \sum_{j\in O\setminus I} \rho_{I\to O\setminus j}\,\frac{\mathrm{vol}(O\setminus j,j)}{\mathrm{vol}(O\setminus j,(O\setminus j)^c)}. \tag{1.3}$$

In the forward programming (1.2), j effectively iterates over the boundary of I in O, as $\mathrm{vol}(I,j) = 0$ if j is outside that boundary. Therefore, the running time of the forward programming benefits from the sparsity of the network. Unlike the forward programming, the iteration over j in (1.3) cannot be restricted to a smaller set. A corollary of Proposition 1 is that the transition probabilities $\rho_{I\to J}$ are not affected by the rate and the duration of the infection.

Let us now observe some connection with the *path-based analysis*. A permitted permutation or an infection path starting at a node $\boldsymbol{i}_*$, refers to a permutation σ of nodes with $\sigma_1 = \boldsymbol{i}_*$, and such that σ_{k+1} is connected to at least one node in $\{\sigma_1,\dots,\sigma_k\}$, for all $k \in [|\sigma|-1]$. Notice that the probability of observing a given infection path is

$$\mathbb{P}\Big(\text{path } \sigma \text{ observed} \mid \sigma_1 = \boldsymbol{i}_*\Big) = \prod_{k=1}^{|\sigma|-1} \frac{\mathrm{vol}(\sigma_{[k]},\sigma_{k+1})}{\mathrm{vol}(\sigma_{[k]},\sigma_{[k]}^c)} \tag{1.4}$$

where $\sigma_I := (\sigma_i \mid i \in I)$. One can obtain the transition probability $\rho_{\{\boldsymbol{i}_*\} \to O}$ by summing (1.4) over all infection paths σ such that $\sigma_1 = \boldsymbol{i}_*$ and $\{\sigma_1, \ldots, \sigma_k\} = O$. Our recursive representation is novel, avoids these explicit summations, and will be key in deriving approximation schemes for $\rho_{I \to O}$ in Section 1.4.

Path-based approaches such as Jordan center [ZY16a] forgo computing the complete likelihood (i.e., avoid summing the odds of all infection paths) and instead find the most probable path, that is, one that maximizes (1.4) in a spanning tree. In contrast, equations (1.2) and (1.3) compute the complete likelihood of the infection set, which has the following advantages over the path-based likelihood: It fully exploits the structure of the graph inside the infection set, not just a spanning tree or a permitted permutation of nodes in the infected subgraph. Moreover, it takes into account the boundary of the infected subgraph via $\operatorname{vol}(I, I^c)$.

Remark 2. Some previous papers, such as [LMO14, ALS15], considered the discrete-time susceptible-infected dynamic. In that setup, the rate and time parameters are intertwined with the transition probabilities in a way that it is hard or infeasible to disentangle them. Therefore, the authors proposed to take β and t as inputs or estimate the probabilities for multiple candidates for the infection time. In this sense, our approach studies a more realistic model with less adverse consequences for estimation.

1.4 Approximations

We now provide two approximations to the likelihood function $L_O(I)$ based on the exact dynamic programming developed in Proposition 1.

Greedy Elimination (GE). We can obtain a singleton source set $I = \{i\}$ that maximizes $\rho_{I \to O}$ with greedy elimination of elements in O. The algorithm we propose is based on the backward recursion (1.3) and is detailed in Algorithm 1. We start with $O_0 := O$ and consider all maximal proper subsets of O_0 that induce a connected subgraph of G. Among those, we choose the one that maximizes the transition probability to O_0, i.e. $\rho_{O_0 \setminus j \to O_0} = \operatorname{vol}(O_0 \setminus j, j) / \operatorname{vol}(O_0 \setminus j, (O_0 \setminus j)^c)$. Suppose that $O_1 := O_0 \setminus j^*$ is the maximizer. Next, we

Algorithm 1 Greedy Elimination

Input: Graph $G([n], E)$ and $O \subset [n]$.

Output: $\boldsymbol{i}^*_{\mathrm{GE}} \in O$.

$O_0 := O$

for $i := 0$ to $|O| - 2$ **do**

  $O'_i := \{j \in O_i : G_{O_i \setminus j}$ remains connected$\}$

  $j^* := \mathrm{argmax}_{j \in O'_i} \dfrac{\mathrm{vol}(O_i \setminus j, j)}{\mathrm{vol}(O_i \setminus j, (O_i \setminus j)^c)}$.

  $O_{i+1} := O_i \setminus j^*$.

end for

$\boldsymbol{i}^*_{\mathrm{GE}} :=$ the single element in $O_{|O|-1}$.

iterate the same procedure for O_1 and so forth, until we reach a singleton set $I := O_{|O|-1}$. The procedure has an $O(k^2 m)$ runtime where $k = |O|$and m is the number of edges in the infected subgraph, G_O.

GE has a Bayesian justification. Let $\widetilde{O}_k$ be the random infected set after k steps. Suppose that we want to find the MAP for $\widetilde{O}_{k-1}$ given $\widetilde{O}_k$. The Bayesian posterior probability is

$$\mathbb{P}(\widetilde{O}_{k-1} = O \setminus j \,|\, \widetilde{O}_k = O) \;\propto\; \rho_{O \setminus j \to O} \cdot P(\widetilde{O}_{k-1} = O \setminus j).$$

Whenever $G_{O \setminus j}$ is connected, the prior is positive. GE finds a proxy for MAP through maximizing the evidence and ensuring the prior is positive.

Algorithm 1 has similarities with finding the most likely path from a source to the observed snapshot. Chang et. al. [CZC15a] propose a similar path-based search called GSBA. They start from each node in O and approximate the most likely path and use it as a proxy to the most likely source. Algorithm 1, however, does this greedy search in a backward fashion.

Mean-field Approximation (MFA). We now approximate $\rho_{I \to O}$ by the mean-field technique. The idea is to treat the set function $I \mapsto \rho_{I \to O}$ as if it was a distribution (or measure) on O and approximate it by the product of its marginals. Fix a subset $O \subset [n]$. For any $I \subset O$, let $\boldsymbol{x}^I = (x^I_j)_{j \in O}$ be the binary representation of I, i.e. $x^I_j = 1\{j \in I\}$ for any $j \in O$.

Algorithm 2 Mean-Field Approximation

Input: Graph $G([n], E)$ and $O \subset [n]$.

Output: $\boldsymbol{i}^*_{\mathrm{MFA}} \in O$.

1: Compute $S, \boldsymbol{z}$ as defined in (1.7).

2: $\widehat{\boldsymbol{b}} := S^{-1}\boldsymbol{z}$.

3: $\boldsymbol{i}^*_{\mathrm{MFA}} := \mathrm{argmax}_{j \in O}\, \widehat{\boldsymbol{b}}_j$.

We find α_0 and $(b_j)_{j \in O}$ such that

$$\widehat{\rho}_{I \to O} = \alpha_0 \prod_{j \in O} b_j^{x_j^I - 1} \tag{1.5}$$

is a good approximation to $\rho_{I \to O}$ for all $I \subset O$, in the sense of minimizing the quadratic deviation from the solution of the recursion (1.2). First note that $\alpha_0 = 1$ since $\rho_{O \to O} = 1$. Next, we plug-in $\widehat{\rho}_{I \to O}$ into the forward recursion, to get

$$\mathrm{vol}(I, I^c)\, \widehat{\rho}_{I \to O} - \sum_{j \in O \setminus I} \mathrm{vol}(I, j)\, \widehat{\rho}_{I \cup \{j\} \to O} = 0.$$

Dividing both sides by $\prod_{j \in O \setminus I} b_j$ gives $\mathrm{vol}(I, I^c) - \sum_{j \in O \setminus I} \mathrm{vol}(I, j)\, b_j = 0$. These equations in general cannot be satisfied exactly for all $I \subset O$. Instead, letting $\boldsymbol{b} = (b_j)_{j \in O}$, we solve the following least-squares problem:

$$\widehat{\boldsymbol{b}} \;\in\; \operatorname*{argmin}_{\boldsymbol{b}} \sum_{I:\, I \subset O} \Big(\mathrm{vol}(I, I^c) - \sum_{j \in O \setminus I} \mathrm{vol}(I, j)\, b_j \Big)^2 = \operatorname*{argmin}_{\boldsymbol{b}} \|Q\boldsymbol{b} - \boldsymbol{r}\|_2^2 \tag{1.6}$$

where $Q \in \mathbb{R}^{(2^{|O|}-1) \times |O|}$ and $\boldsymbol{r} \in \mathbb{R}^{(2^{|O|}-1) \times 1}$ are defined as follows:

$$Q_{I,j} = 1\{j \notin I\}\ \mathrm{vol}(I, j),\ \forall\, I \subset O,\ j \in O, \qquad \boldsymbol{r}_I = \mathrm{vol}(I, I^c),\ \forall\, I \subset O.$$

The solution of (1.6) satisfies the normal equations $Q^T Q \widehat{\boldsymbol{b}} = Q^T \boldsymbol{r}$. The following proposition shows that $Q^T Q$ and $Q^T \boldsymbol{r}$ can be computed efficiently. Let A be the adjacency matrix of the network.

Proposition 2. *The solution* $\widehat{\boldsymbol{b}}$ *of* (1.6) *satisfies the linear system* $S\widehat{\boldsymbol{b}} = \boldsymbol{z}$ *with* S *and* $\boldsymbol{z}$

given by

$$
\begin{aligned}
S = \Xi\Big(& A_{OO} \odot A_{OO}^T + A_{OO}^T A_{OO} - A_{OO} \odot (\boldsymbol{u}\mathbf{1}^T) \\
& - A_{OO}^T \odot (\mathbf{1}\boldsymbol{u}^T) + \boldsymbol{u}\boldsymbol{u}^T \Big) \in \mathbb{R}^{|O|\times|O|}, \\
\boldsymbol{z} = & (\mathbf{1}^T\boldsymbol{u} + 2\mathbf{1}^T\boldsymbol{v})\boldsymbol{u} - 2\boldsymbol{v} \odot \boldsymbol{u} + 2A_{OO}\,\boldsymbol{v} + (\boldsymbol{u} - \boldsymbol{u}^{out}) \odot \boldsymbol{u} \\
& + \Big((A_{OO} + A_{OO}^T) \odot A_{OO}^T\Big)\mathbf{1} + A_{OO}^T(\boldsymbol{u}^{out} - \boldsymbol{u})
\end{aligned}
\tag{1.7}
$$

where $\boldsymbol{u} := A_{OO}^T\mathbf{1}$, $\boldsymbol{u}^{out} := A_{OO}\mathbf{1}$, *and* $\boldsymbol{v} = A_{OO^c}\mathbf{1}$. *Here* $\odot$ *is the element-wise matrix product,* $\Xi(\cdot)$ *is a matrix operator that returns the same matrix with double the diagonal entries, and* $\mathbf{1}$ *is the vector of all ones.*

See Appendix 4.2.2 for the proof. Proposition 2 shows that the mean-field approach reduces to solving a linear system of equations in $|O|$ variables, a task with much better computational complexity than solving the original recursion. Both S and $\boldsymbol{z}$ can be computed in at most $O(k^2)$ time, where $k = |O|$. In the cases where A is sparse (which often the case for real networks), S will be a rank-one perturbation of a sparse matrix (both A_{OO} and $A_{OO}^T A_{OO}$ will be sparse), hence solving the resulting system is often much faster than the worst-case, i.e., faster than $O(k^3)$.

Remark 3. MFA and GE utilize the forward and backward programs ((1.2),(1.3)), respectively. We have tried to apply linearization to the backward program and greedy inclusion to the forward program. However, the former does not go through as smoothly and the latter leads to a sub-par method. Whether one can utilize both recursions simultaneously to achieve a better performance is open.

1.5 Simulations

The methods proposed in this paper, the Greedy Elimination (GE) and the Mean Field Approximation (MFA), show superior performance in source identification, compared to popular procedures, while having comparable runtimes. In this section, we make a comparison based on these two measures (source identification ability and runtime) on real and synthetic

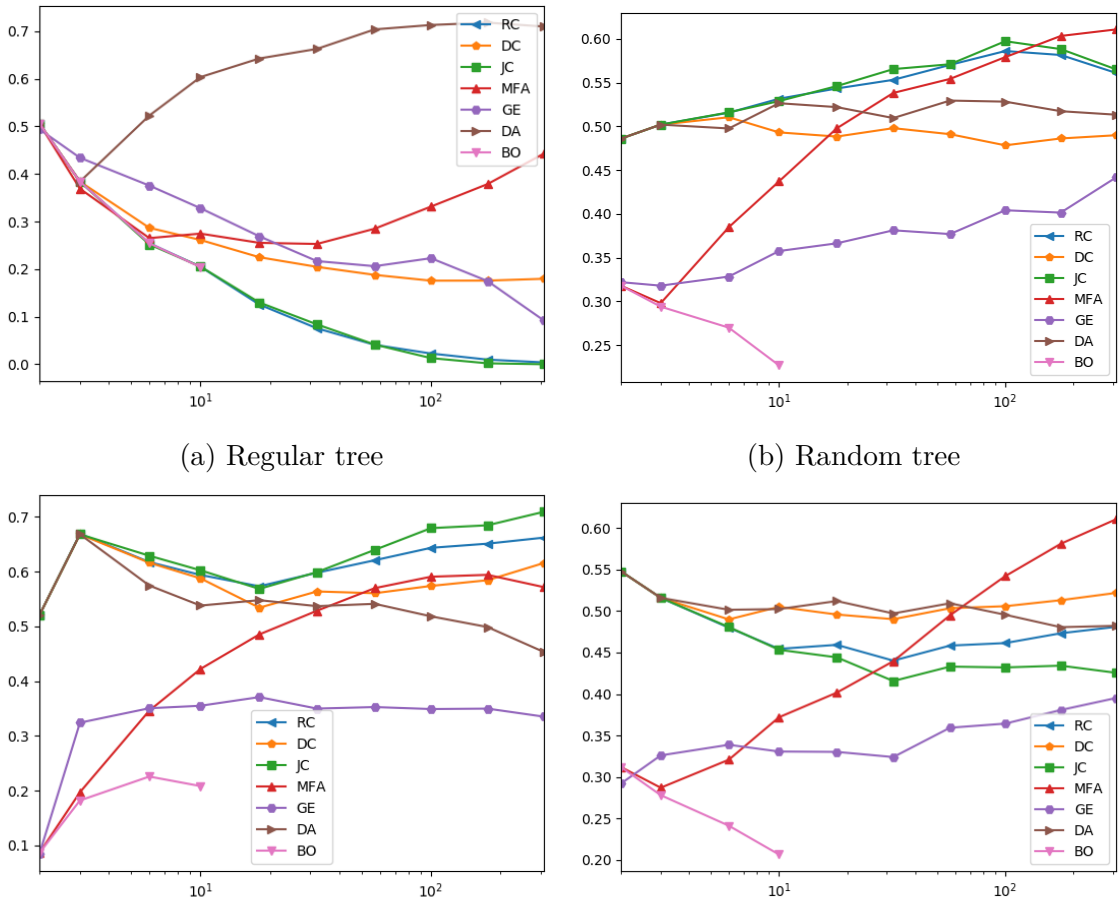
Table 1.1: Network statistics

Network	Internet	Power	Wiki vote	UCSC68	UC64	DC-SBM
n	10670	4941	7066	8979	6810	1962
Mean degree	4	3	29	50	46	66
Max. degree	2312	19	1065	454	660	897
Clust. coeff.	0.01	0.10	0.13	0.17	0.19	0.30

networks. As discussed in Section 1.2.2, we consider ranking estimators (i.e., those that output a permutation of the nodes according to their likelihood of being the source) and focus on the rank loss. If the method does not return a ranking, we tweak it to do so. We evaluate the methods based on the expected rank, $\mathbb{E}[R]$, where R is the rank of the true source among the list of candidates (cf. Section 1.2.2). The expectation is taken with respect to the variation in choosing the true source, which is drawn at random from the entire network. We normalize the expected rank to get a number in $[0, 1]$, with zero corresponding to perfect recovery, i.e., we use $(\mathbb{E}[R] - 1)/n$.

We consider a variety of real and simulated networks. Our selection includes an Internet Autonomous System [Ore, LKF05], US west-coast power grid [WS98], two Facebook-100 networks [TKM11, TMP12], called UC64 and UCSC68, and a Wikipedia voting network [LHK10]. In addition, we present our results on a number of synthetic networks that are well studied in the literature, including regular trees, random trees, and degree-corrected stochastic block models (DC-SBM) [KN11].

Table 1.1 summarizes the statistics on the largest connected component of these networks. The regular tree is of degree 3 and depth 10. The random tree has 500 nodes. For the DC-SBM network, we generate from a 3-community planted partition version, i.e., $\mathbb{E}[A_{ij}] = \theta_i \theta_j P_{ij}$ where $P_{ij} = 0.5$ if nodes i and j are in the same community and $P_{ij} = 0.02$ if they are in different communities. The degree parameters θ_i are generated from a rescaled Pareto distribution with $\alpha = 2$ and threshold $= 1$.



(a) Regular tree

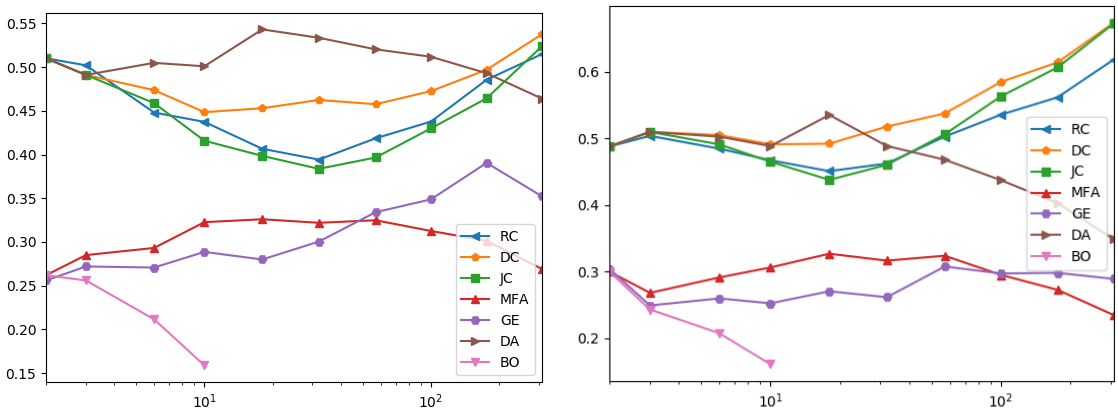
(b) Random tree

(c) Internet AS

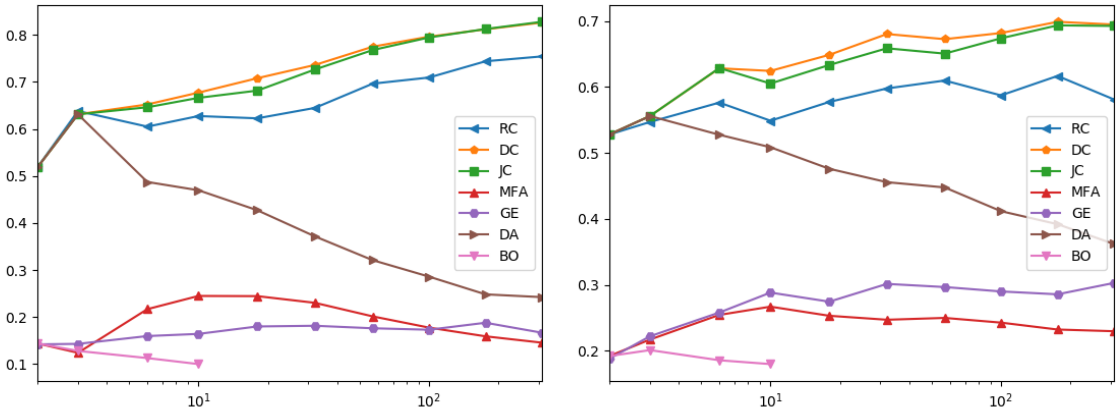
(d) US West Power Grid

Figure 1.1: Plots of the expected relative rank versus the infection size for low-transitivity networks.

The results are illustrated in Figures 3.2, 1.2. The methods we consider besides the optimal Bayes solution (BO), the MFA, and the GE are the Rumor Centrality (RC), the Degree Centrality (DC), the Jordan Center (JC) and the Dynamical Age (DA). Our selection of the methods loosely follows the methods surveyed in [JWY17]. Each curve shows the performance of one method for different values of the infection size, $2 \leq |O| \leq 300$. Each point is an average over 500 infection paths rooted at random sources. To avoid an unreasonable computation time, we skip the BO for the infected sets of size greater than 10. The BO curve serves as the benchmark for the best achievable performance. Note that even the optimal solution needs to output a large set to catch the source, signifying the inherent difficulty of



(a) UCSC68 (b) UC64



(c) Wiki vote (d) DC-SBM

Figure 1.2: Plots of the expected relative rank versus the infection size for high-transitivity networks.

the problem.

Rumor and Jordan centralities perform optimally on regular trees in Figure 1.1a, as predicted by the theory [SZ11a, ZY16a], although the network here is not exactly an *infinite* tree. Notice that RC, JC, and BO overlap for infection sizes not exceeding the depth of the tree. Degree centrality also turns out to be a close competitor in this figure. Moving to other networks, however, these popular methods do not perform better than random guessing. For all the three, the expected relative rank is close to 0.5, even in a random tree. The plots in this section show that, despite their popularity, the RC and JC are quite unreliable for source recovery.

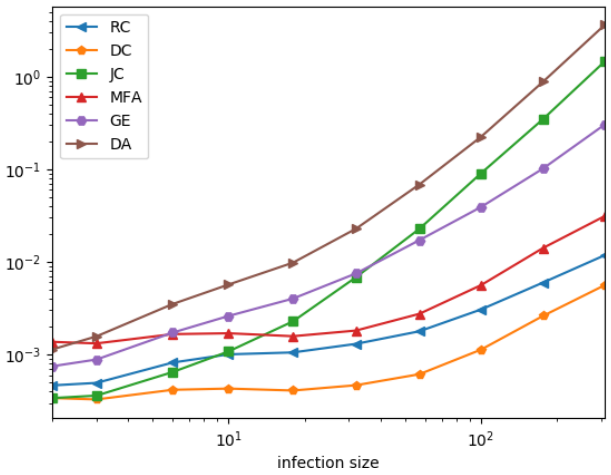


Figure 1.3: Runtime in seconds.

DA tends to perform well only when the infection size is sufficiently large. In some of the networks, i.e., in Figures 1.2b, 1.2c, and 1.2d, it is a close competitor to GE and MFA, while still behind them with a margin. DA also performs close to GE in the "Internet AS" network (Figure 1.1c).

Among our proposed methods, MFA outperforms RC, JC, DC, and DA in Figures 1.2a, 1.2b, 1.2c, 1.2d. MFA ranks the true source, on average, in the top 30%. The networks with superior MFA performance have the highest transitivity (a.k.a. clustering coefficient) in Table 1.1, that is, many triangles among triples of nodes. Transitivity has been studied extensively and distinguishes human social networks from random trees and less cyclical networks, such as the water distribution systems and traffic networks. In this sense, MFA is suitable for rumor source detection in social networks.

GE is the global winner, except in regular trees (Figure 1.1a). We were surprised that a greedy algorithm had such a widespread success. GE not only performs well in highly transitive networks, but also outperforms RC, JC, DC, and DA on random trees (Figure 1.1b) and less transitive networks (Figures 1.1c, 1.1d).

Figure 1.3 illustrates the runtimes (on the log scale) for a single run on the UC64 network. Degree centrality is the fastest, followed by RC, MFA, GE, JC, and DA. The first three have comparable speed and scale quite well. In contrast, although the runtime for JC starts as low

as that of RC, it accelerates past GE as the infection size grows. DA and JC do not scale well and GE follows them by a margin. It is worth noting that [LTL14] gives a linear-time implementation of JC on trees that we have not tested here. BO is removed from this plot since its runtime grows exponentially with the infection size.

Based on these results, we advocate for the use of GE as the main tool for identifying sources of epidemics, regardless of the network topology or the nature of the epidemics (rumor propagation, disease contagion, etc.). MFA should be applied with caution. It is superior in social (transitive) networks, and attractive for its simplicity, scalability, and the potential for parallelism.

CHAPTER 2

Monte Carlo Estimation

2.1 Introduction

In this chapter, we study a class of epidemic inference problems that aim to estimate the parameters of the propagation dynamics from data, and more specifically from a static snapshot of the epidemics. We propose a unifying formulation for this class of problems and introduce an efficient Importance Sampling (IS) solution that can be applied in general.

The IS approach can be used to efficiently solve many types of epidemic inference problems including: 1) retrieving the parameters of the epidemic, such as the rate, the duration and the source of the spread, 2) assessing whether the spread is contagious or random, and 3) reconstructing the epidemic network. To keep this chapter focused, however, we narrow down our attention to one of the applications, namely, the epidemic source recovery problem. With the observation of the full infected set of nodes in a network, the task is to identify the source, also known as patient zero. This problem has been studied extensively but the solutions so far have not been satisfactory [JWY16]. The main shortcoming in this literature is that the current solutions do not perform well under real world complex networks [SZ11b, FCP14, ZY16b] or turn out to be prohibitively slow as the infection size grows [LMO14, CZC15b, ABD14]. We offer a fast method, with strong theoretical guarantees and show that it performs well in practice on real complex networks. Here is a summary of the contributions of this paper:

- We propose a novel importance sampling method, for computing the probability of the infection set under a continuous heterogeneous susceptible-infected dynamics and show it provides an efficient solution to the source recovery problem.

- We provide an error bound for the quality of approximation obtained by this method and show that a satisfactory precision can be achieved in polynomial-time.
- We compare our approach, under extensive simulations, with peer methods including previous Monte Carlo approaches. The conclusion is that the proposed importance sampling is much more efficient (i.e., requires much less samples to achieve a given accuracy).

In what follows, we define the setup, including the continuous heterogeneous susceptible-infected dynamics and argue for its generality. Then, the Bayes optimal solution to the source recovery problem is formalized. We present our importance sampling method afterwards, following a discussion of the previous Monte Carlo approaches. We then develop a theoretical bound on the relative mean-square error of the method. The paper concludes with a presentation of simulation results illustrating the performance of our method applied to multiple real and synthetic networks.

The Epidemic Model

Numerous attempts at explaining network epidemics has led to four main categories of information spread models: Susceptible-infected (SI) type models, the Information Cascade (IC) model, the Linear Threshold Model (LTM), and the Game Theoretic Model (GTM) [LWG17]. Among these, the SI dynamics and its variants have been used extensively in studying epidemics [KMS17] and have gained significant popularity. The model originates from population ODEs that describe the evolution of the infected population size [KM27]. It has been extended to Markov processes on networks that rule the spread of a disease (or social habit, information, etc.) from infected nodes to susceptible neighbors.

Some variations of the SI model include extra states in addition to being susceptible or infected. The most widely used are the Susceptible-Infected-Recovered (SIR) model, which allows for a recovery state after the infection, the Susceptible-Infected-Susceptible (SIS) model, where an infected node will recover without immunization, the SEIR (E stands for exposed),

which adds a state of being infected but not infectious. Graph-coupled Hidden Markov Model (GCHMM) [DPH12] is another recent example, which follows a SIS (with external noise) model but takes the contagious infection as a hidden state. Some extensions introduce variations in the rate of the spread between neighboring nodes. The FSIR model [FHL15], for instance, assumes the rate of information spread is conversely proportional to the degree of a node, so that hubs become less prone to notice and retweet a message from a friend (due to limited attention). The ESIS model [WLJ15], on the other hand, introduces weights to the information spread according the emotional ties between friends and the sentiment of the message. See [LWG17, KMS17] for additional variations and extensions of the SI model.

2.1.1 The CH-SI dynamic

In this paper, we assume that the epidemic is governed by a continuous heterogeneous susceptible-infected (CH-SI) dynamic. Let $G(V, E)$ be a connected, directed, and weighted graph defined by the set of vertices $V = [n]$ and the set of edges E, with adjacency matrix $A = (a_{ij})_{ij}$ representing the edge weights. Each node at continuous time t can be in an infected, $x_i(t) = 1$, or susceptible state, $x_i(t) = 0$. A susceptible node $i \in V$ is in risk of getting infected at the exponential rate $\lambda \sum_{j \in V} A_{ji} x_j(t)$, independent of other nodes. Here, $\lambda > 0$ determines the rate of spread. Infection is an absorbing state for all nodes

At time $t = 0$, a node s is picked at random, according to distribution $p(\cdot)$, to have $x_s(0) = 1$. All others remain susceptible. As the infection grows, it finally covers the entire network. Let $\widetilde{\sigma} \in \mathrm{Sym}(V)$ denote the permutation of the nodes in the order they get infected. We refer to $\widetilde{\sigma}$ as the *infection path*. Here, $\mathrm{Sym}(V)$ refers to the set of permutations of V. We use $\widetilde{\sigma}_i$ to denote the i-th infected node and $\widetilde{\sigma}_{[i]}$ to denote the infection path truncated at its i-th element. The infection path does not reveal any information about the timing of infection. We write

$$\widetilde{\sigma} \sim \text{CH-SIP}(G, s) \tag{2.1}$$

to denote the probability distribution for $\widetilde{\sigma}$, given $\widetilde{\sigma}_1 = s$, that is, the source of infection is $s \in V$. CH-SIP stands for continuous (heterogeneous) susceptible-infected path. In the next

section, we will characterize CH-SIP and show that it is independent of λ.

Let $\widetilde{O}_i$ be the set of the first i infected nodes and call it the *infection set* or *snapshot* of size i. In other words,

$$\widetilde{O}_i = \{\widetilde{\sigma}_j \,:\, 1 \leq j \leq i\} \tag{2.2}$$

The event that $\{\widetilde{O}_i = O\}$ implies that all $j \in O$ are infected prior to any $j \in V \backslash O$. The infection set hides the information about the order in which nodes are infected. We define a *cascade* of snapshots as a series of snapshots $(\widetilde{O}_{i_1}, \ldots, \widetilde{O}_{i_m})$ where $1 \leq i_1 < \cdots < i_m \leq n$. Obviously, for a cascade, we have $\widetilde{O}_{i_1} \subset \cdots \subset \widetilde{O}_{i_m}$. The event that $(O_1, \ldots, O_m)$ is a cascade of infection sets implies that the nodes in O_i are infected before the nodes in O_{i+1} for all $i \in [m-1]$ and the nodes in O_m are infected prior to the rest of the nodes in V.

In a network which is not instantaneously monitored, what we can observe from an epidemic is a single (or a cascade of) snapshot(s). In the next section, we will discuss the maximum likelihood approach to inferring parameters of an epidemic and/or the underlying network G from a single or a cascade of snapshots.

2.1.2 Flexibility of the model

Before proceeding to the details of the methods, let us pause for a moment to discuss the generality of the continuous heterogeneous susceptible infected dynamic. CH-SI is a flexible extension of the SI dynamic that captures the variety in the rate at which the information/disease is transmitted from one neighbor to another. Examples of various factors that cause this distinction across the edges of the network are: the distances between cities, the length of water pipes or electric wires, the strength of emotional ties between friends, the limited attention span due to the abundance of friends, different levels of immunity to an infection among citizens, and so on. In this respect, CH-SI generalizes the FSI and ESI, i.e., the FSIR and ESIR without the recovery state, and allows one to incorporate node and edge attributes in characterizing the diffusion. We will not discuss how these attributes are linked to the speed of the information transmission. Instead, will assume that the transmission rates are factored into the edge weights up to a scalar.

We abide by the continuous nature of the original SI dynamic in this paper. Many prior works mentioned in the introduction adopt a discrete version of SI. In this respect, we solve the epidemic inference problem in a more general and realistic setup. As we will see in the next section, arguing in infinitesimal time steps, the solution can be relaxed to one degree of freedom, since the time parameter, i.e., when we happen to observe the snapshot, will disappear from the equations and is no longer required to be known a priori (or otherwise estimated).

A drawback of CH-SI is the absence of recovery state. In many real situations, it is natural to think of individuals as developing immunization to a disease, abandoning a habit, or removing a tweet from their account. While it is beneficial to add the recovery state, it is conceivable that in a short period after the outbreak, no one develops immunization or abandons propagating the information. We hope in future we can extend the methods developed here to an epidemics with recovery state.

2.2 Epidemic Inference from Complete Snapshots

We now formulate a class of epidemic inference problems that can be efficiently solved by our sampling approach. Suppose that an epidemic outbreak starts from node $s \in V$, in network G, and propagates according to a CH-SI dynamic, with a rate controlled by λ and A, the weighted (asymmetric) adjacency of G. At some point, we observe the full snapshot of the infected nodes to be $O \subset V$ (or we observe a cascade of snapshots at different points in time). We would like to exploit this knowledge to infer either of the following unknowns:

1. *Epidemic source identification*: Given access to the links and their strength, E or equivalently A, which node is more likely to be the source, s [SZ11b]?

2. *Testing hypotheses about the epidemic network*: given two candidates for the contact network, $G_1(V, E_1)$ and $G_2(V, E_2)$, which one is more likely to be the underlying network [MCM15, KL17b]?

3. *Epidemic network reconstruction*: given a set of node and edge attributes associated

with G and a parametric association rule between these attributes and the weight matrix A, what are the most likely weights for the infection graph [GLK12, KL18, AAR10]?

These problems have been studied extensively under various epidemic models and different network topologies. Their uniting feature is that their solutions rely on the efficient computation of the likelihood of a single snapshot or a cascade of them. As shown in this section (although not proved here), the likelihood is a partition function with exponential number of terms and the exact solution is not feasible. See [NGM16b] for the complexity lower bound in a particular version of the source recovery problem. The earlier approaches appealed to heuristics [ZY16b, FCP14, CZC15b], variational methods [LMO14, ABD14, KA19, ABD14], and Monte Carlo techniques [ALS15] to approximate the likelihood. For the rest of this paper, we restrict our attention to epidemic source recovery problem (problem 1 in the list above). The analysis is more or less applicable to other problems.

We begin by defining the probability of the event that $O \subset V$ is a snapshot of the epidemic. Suppose $k = |O|$, and consider a "candidate s" for the source. Let us write

$$\mathbb{P}_s(\cdot) \equiv \mathbb{P}(\,\cdot \mid \widetilde{\sigma}_1 = s)$$

to denote the distribution (2.1) of a path started at s. Then,

$$\rho(s \to O) := \mathbb{P}_s(\widetilde{O}_k = O) = \sum_{\sigma \in \mathrm{Sym}(O)} \mathbb{P}_s(\widetilde{\sigma}_{[k]} = \sigma) \tag{2.3}$$

is the probability of transition from s to O. The sum on the RHS potentially contains $\Theta(|\mathrm{Sym}(O)|)$ nonzero terms. Even if G_O is a tree, the complexity of computing this formula can be non-polynomial in k. The difficulty of solving epidemic inference problems listed earlier is due to the complexity of evaluating this sum for a general network topology.

Let us introduce some notation: Let $\mathrm{vol}(U, W) := \sum_{u \in U, w \in W} A_{uw}$ be the the weight of the cut between nodes U and W in the network. We also write

$$\partial\mathrm{vol}(U, W) := \mathrm{vol}(U, W \setminus U).$$

For singleton sets $\{i\}$, we often drop the braces and write i.

As was previously shown [CZC15b, KA19, KL17b], the odds of a certain infection path governed by the SI dynamic has a time and rate-free form and can be computed in linear time. The same applies to the CH-SIP p.m.f.,

$$\begin{aligned}\mathbb{P}_s(\widetilde{\sigma}_{[k]} = \sigma_{[k]}) &= \prod_{i=0}^{k-1} \mathbb{P}_s(\widetilde{\sigma}_{i+1} = \sigma_{i+1} \mid \widetilde{\sigma}_{[i]} = \sigma_{[i]}) \\ &= 1_{\{\sigma_1 = s\}} \prod_{i=1}^{k-1} \frac{\partial \mathrm{vol}(\sigma_{[i]}, \sigma_{[i+1]})}{\partial \mathrm{vol}(\sigma_{[i]}, [n])} \end{aligned} \tag{2.4}$$

for all $k \in [n]$, with the convention $[0] = \emptyset$. Note that $\partial \mathrm{vol}(\sigma_{[i]}, \sigma_{[i+1]}) = \mathrm{vol}(\sigma_{[i]}, \sigma_{i+1})$. This equation follows easily using the Markov property of the dynamics and a classic property regarding the minimum of independent exponential random variables [Ros14]. Note that (2.4) does not depend on either the rate λ or the time of observation.

With the complete snapshot $\widetilde{O}_k = O$ as the observation, the maximum likelihood estimator for the epidemic source follows:

$$\widehat{s}_{\mathrm{ML}}(O) := \operatorname*{argmin}_s \ell(s \to O). \tag{2.5}$$

where

$$ew\ell(s \to O) = -\log \rho(s \to O) = -\log \mathbb{P}_s(\widetilde{O}_k = O) \tag{2.6}$$

is the negative log-likelihood. This estimator also coincides with the Bayes optimal estimator for the "zero-one" or "rank" loss and a uniform prior on the source. In the case of the rank loss, the estimator is the entire ranking obtained by sorting $s \mapsto \ell(s \to O)$ to get an increasing sequence.

2.2.1 Evaluation

We consider estimators for the source recovery problem that generate a ranking on the candidates. Consider an epidemic source estimator that ranks the elements of O according to permutation $\hat{\pi} : O \to [k]$, where $\hat{\pi}^{-1}(1)$ is the most likely candidate for the source, $\hat{\pi}^{-1}(2)$ the second most likely candidate and so on.

We measure the performance of $\hat{\pi}$ by how it ranks the optimal Bayes source estimate (assuming a uniform prior). Consider the normalized rank of the optimal Bayes estimate,

$\hat{s}_{\mathrm{ML}}(O)$, in $\hat{\pi}$, that is,

$$R_B(\hat{\pi}) := \frac{1}{k} \sum_{s \in O} 1\{\hat{\pi}(\hat{s}_{\mathrm{ML}}) > \hat{\pi}(s)\}.$$

We refer to R_B as the *Bayes rank* which is a measure of how close a ranking is to the Bayes optimal estimator. (Note that Bayes rank is a loss defined between a ranking and a candidate source, not between two rankings. The loss does not penalize disagreements with Bayes optimal ranking below its top element. This makes it a more appropriate measure for the source recovery problem.) Instead of working with R_B, we introduce a more natural metric that we call *weighted Bayes rank*,

$$\mathcal{R}_B(\hat{\pi}) := \frac{1}{k} \sum_{s \in O} \Big[1 - \frac{\ell(\hat{s}_{\mathrm{ML}} \to O)}{\ell(s \to O)}\Big] 1\Big\{\hat{\pi}(\hat{s}_{\mathrm{ML}}) > \hat{\pi}(s)\Big\}. \tag{2.7}$$

This measure is less sensitive to the disagreements between pairs with similar posterior probabilities. This variation to the Bayes rank is inspired by the weighted version of the Kemeny distance (also known as Kendall's τ) introduced in [NOS12] for the pairwise ranking problem.

2.3 Monte Carlo Estimation

We now derive Monte Carlo estimates for $\mathbb{P}_s(\widetilde{O}_k = O)$ for any $O \subset V$ and any $s \in O$. The first two methods we introduce here are the *Direct Monte Carlo* (DMC) and *Soft-Margin Monte Carlo* (SMC), which are developed by [ALS15], for discrete susceptible-infected dynamics. We redefine them for the continuous case and provide a faster implementation using the parameter-free formula for CH-SIP, developed in (2.4). Additionally, we will provide lower bounds for the sampling error of DMC to demonstrate the inefficiency of this method in large graphs. In the simulations, the performance of SMC will be assessed empirically. We end the section with our main contribution in this paper, the Importance Sampling (IS), and will prove a guarantee for its fast convergence to the likelihood.

2.3.1 Direct Monte Carlo sampling

Let us draw N infection paths from CH-SIP rooted at s:

$$\widetilde{\sigma}^t \sim \text{CH-SIP}(G, s),\ t = 1, \ldots, N. \tag{2.8}$$

The direct Monte Carlo estimate is then

$$\widehat{\rho}_{\text{MC}}(s \to O) = \frac{1}{N} \sum_{t=1}^{N} 1\{\widetilde{\sigma}^t_{[k]} \subset O\}. \tag{2.9}$$

We sample the infection path incrementally according to (2.4), i.e., we draw the $(i+1)$th infected point according to

$$\mathbb{P}_s(\widetilde{\sigma}_{i+1} = v \mid \widetilde{\sigma}_{[i]}) = \frac{\text{vol}(\sigma_{[i]}, v)}{\partial \text{vol}(\sigma_{[i]}, [n])},\ v \in \partial \widetilde{\sigma}_{[i]} \tag{2.10}$$

where $\partial U := \{u \in V \mid A_{iu} > 0\ \exists i \in U\}$. We may stop the progress at step k, the size of the observed infection set, when we reach $\widetilde{\sigma}_{[k]}$. However, since we can evaluate the event $1\{\widetilde{\sigma}^t_{[k]} \subset O\}$ at the first time a node outside O is infected, we will employ early stopping which can dramatically improve the simulation speed.

Let us now argue why controlling the error of Direct MC may require a non-polynomial number of draws. First, note that DMC is an unbiased estimator, since

$$\mathbb{E}_s 1\{\widetilde{\sigma}^t_{[k]} \subset O\} = \mathbb{P}_s(\widetilde{O}_k = O). \tag{2.11}$$

To investigate the typical behavior of DMC, we consider the following random graph model

Definition 1. We say that G is an inhomogeneous (directed) Erdös-Rényi graph with edge probability matrix $P = (P_{ij}) \in [0,1]^{n \times n}$, denoted as $G \sim \text{ER}(n, P)$, if $A_{ij} \sim \text{Ber}(P_{ij})$ independently for all $1 \le i, j \le n$.

We have the following lower-bound on the performance of the DMC sampling:

Theorem 1. *Let $G \sim \text{ER}(n, P)$ as in Definition 1, and that $n \ge 16$. Assume that $0 < p_{\min} \le P_{ij} \le p_{\max}$ for all $1 \le i, j \le n$, and let $\beta = p^2_{\min}/(p_{\min} + 2p_{\max})^2$. Let O be a subset of nodes of size k satisfying $23/p^2_{\min} \le k \le \sqrt{n}$. Take $s \in O$ and let $\rho(s) := \rho(s \to O)$. Then,*

$$\mathbb{P}\Big(\rho(s) \le (\beta n)^{-k/4}\Big) \ge \frac{1}{2}. \tag{2.12}$$

Moreover, let $\widehat{\rho}_{MC}^{(N)}$ *be the DMC estimate of* $\rho(s)$ *based on* N *draws as in* (2.9)*. Then, for any* $N \leq (\beta n)^{k/4}$*,*

$$\mathbb{P}\Big(\widehat{\rho}_{MC}^{(N)} = 0\Big) \geq \frac{1}{4}. \tag{2.13}$$

Theorem 1 shows that the number of runs necessary to bound the error of DMC is exponential in the infection size k and grows with the network size n.

It is interesting to note that in order to establish (2.12), showing that the snapshot probabilities are exponentially small, we will use the unbiasedness of the importance sampler introduced later (see Lemma 1). In other words, the importance sampling idea also provides a theoretical device in understanding the asymptotic behavior of transition probabilities.

2.3.2 Soft-Margin Monte Carlo sampling

Another approach, proposed by [ALS15], is to replace the Bernoulli measure $1\{\widetilde{\sigma}_{[k]}^t \subset O\}$ with a smooth version. Let

$$\widehat{\rho}_{\mathrm{SMC}}(s \to O) = \frac{1}{N}\sum_{t=1}^{N} h \circ \varphi(\widetilde{\sigma}_{[k]}^t, O) \tag{2.14}$$

where $\varphi(U, W) = |U \cap W|/|U \cup W|$ is the Jaccard distance between sets U and W, $h(t) = \exp(-(t-1)^2/a^2)$ and $a > 0$ is a fixed scalar. In [ALS15], the authors did not argue why this smoother version may improve the performance and if the so, to what extent. However, our simulations reconfirm that the smoothing does improve the performance over direct MC sampling in real networks.

2.3.3 Importance sampling

Since $\rho(s \to O)$ is quite small for large k, it is very unlikely that a path generated from the dynamic on the whole network lies entirely within O. This means that directly sampling from the target distribution generates zero terms in (2.9) in the majority of runs, requiring (exponentially) many runs to get an accurate estimate, as formalized by Theorem 1. To avoid this, the idea of importance sampling [Gew89] is to tweak the proposal distribution to avoid

these zero contributions. The scheme we propose is to constrain the dynamics to G_O and draw infection paths accordingly. These paths are by design guaranteed to always hit O, as opposed to those generated by the DMC. As we will show, sampling in this way greatly reduces the number of runs required for accurate estimation.

More specifically, we draw samples $\{\widetilde{\sigma}^t\}$ from the CH-SIP dynamic restricted to the subgraph of G on O, denoted as G_O. That is, we draw

$$\widetilde{\sigma}^t \sim \text{CH-SIP}(G_O, s),\ t = 1, \ldots, N, \tag{2.15}$$

independently and consider the estimator

$$\widehat{Y}_{\text{unb}} = \frac{1}{N}\sum_{t=1}^{N} Y(\widetilde{\sigma}^t), \quad Y(\widetilde{\sigma}^t) := \prod_{i=1}^{k-1} \frac{\partial \text{vol}(\widetilde{\sigma}^t_{[i]}, O)}{\partial \text{vol}(\widetilde{\sigma}^t_{[i]}, [n])}. \tag{2.16}$$

Let $\widetilde{\sigma} \sim \text{CH-SIP}(G_O, s)$ be a generic sample from the restricted dynamic. We denote the distribution of $\widetilde{\sigma}$ as $\mathbb{P}_s^{G_O}$ to emphasize that it is a dynamic restricted to G_O. We have

$$\mathbb{P}_s^{G_O}(\widetilde{\sigma}^k = \sigma^k) = \prod_{i=1}^{k-1} \frac{\partial \text{vol}(\sigma_{[i]}, \sigma_{[i+1]})}{\partial \text{vol}(\sigma_{[i]}, O)}, \tag{2.17}$$

where the indicator is dropped since $s \in O$. The following shows that $\widehat{Y}_{\text{unb}}$ is an unbiased estimator of $\rho(s \to O)$:

Lemma 1. *For any $s \in O \subset [n]$, we have*

$$\mathbb{E}[\widehat{Y}_{unb}] = \mathbb{E}[Y(\widetilde{\sigma})] = \rho(s \to O). \tag{2.18}$$

where $\widetilde{\sigma} \sim \text{CH-SIP}(G_O, s)$.

Our proposed estimator is a modification of $\widehat{Y}_{\text{unb}}$, namely,

$$\widehat{\ell}_{\text{IS}}(s \to O) := \frac{1}{N}\sum_{i=1}^{N} -\log Y(\widetilde{\sigma}^t), \tag{2.19}$$

which is an estimator of $\ell(s \to O)$. That is, instead of estimating the probabilities directly, we estimate their log. This estimator is no longer unbiased. However, it has the advantage of having a much lower variance, hence requires much smaller Monte Carlo sample size (N) to achieve a given accuracy. This is formalized in Theorem 2 below.

2.4 Consistency

We now provide theoretical guarantees for the importance sampling estimator in (2.19) and the associated source estimator. Our main result establishes an upper bound on the relative error of (2.19) for estimating $\ell(s)$, for a wide class of networks. Our results go substantially beyond existing theoretical guarantees in the epidemic source recovery literature which are all confined to (regular) trees [JWY16].

We make the following two assumptions on the underlying network:

(R1) We say that $O \subset [n]$ has a *regular boundary* (in network G) if there is constant $c_1 > 0$ such that for all $U \subset O$ with $|U| \geq |O|/2 + 1$,

$$\mathrm{vol}(U, O^c) \geq c_1 |U| d_{\mathrm{ave}} \tag{2.20}$$

where d_{ave} is the average degree of O in O^c, that is, $d_{\mathrm{ave}} = \frac{1}{|O|} \sum_{i \in O} \sum_{j \in O^c} A_{ij} = \mathrm{vol}(O, O^c)/|O|$.

(R2) We say that O is *uniformly connected* if there is a constant $c_2 > 0$ such that for any $U' \subset O$,

$$\partial \mathrm{vol}(U', O) \geq c_2. \tag{2.21}$$

For a random variable X, we write $X \sim \mathrm{subG}(\sigma^2)$ if X is sub-Gaussian, with squared sub-Gaussian parameter σ^2, that is, $\mathbb{E}[e^{\lambda(X - \mathbb{E}X)}] \leq e^{\lambda^2 \sigma^2/2}$ for all $\lambda \in \mathbb{R}$. We refer to ? for the background on sub-Gaussian variables. Recall that for such a variable, $\mathrm{var}(X) \lesssim \sigma^2$.

Theorem 2. *Fix $O \subset [n]$ satisfying (R1) and (R2) and let $k = |O|$. Take $\tilde{\sigma}^t, t = 1, \ldots, N$ to be a random sample of paths from the* CH-SIP *dynamic restricted to O, as in* (2.15)*. Let*

$$Z(s \to O) := \frac{\widehat{\ell}_{IS}(s \to O)}{\mathbb{E}\big[\widehat{\ell}_{IS}(s \to O)\big]} - 1$$

be the relative distance of $\widehat{\ell}_{IS}(s \to O)$ from its mean. Then, for any $s \in O$,

$$Z(s \to O) \sim \mathrm{subG}\left(\frac{\log^2(k d_{\max}/c_2)}{\alpha^2 N}\right) \tag{2.22}$$

where $\alpha = \log(1 + c_1 d_{ave}/d_{\max})$. Here, d_{ave} is as defined in (R1) and $d_{\max}$ is the maximum degree of the entire network G.

The inequality in (2.22) indirectly bounds the variance of $\widehat{\ell}_{\mathrm{IS}}(s \to O)$. Theorem 2 thus shows that the number of iterations N necessary to achieve a given accuracy grows at most poly-logarithmically in the infection size k.

As a corollary to Theorem 2, the ranking generated by the IS estimate of the negative log-likelihood, $\widehat{\ell}_{\mathrm{IS}}(\cdot \to O)$, closely follows that of its mean.

Corollary 1. *Let $\hat{\pi}_{IS}$ be the ranking generated by $\widehat{\ell}_{IS}$. Suppose that $\mathbb{E}\,\widehat{\ell}_{IS}$ and ℓ generate identical rankings. Under the assumptions of Theorem 2 and using the same notation, with probability at least $1-\delta$,*

$$\mathcal{R}_B(\hat{\pi}_{IS}) \;\leq\; \sqrt{\frac{2\log^2(kd_{\max}/c_2)\log(1/\delta)}{\alpha^2 N}}.$$

A consequence of Corollary 1 is that for any given $\varepsilon > 0$, taking $N \gtrsim \log^2(kd_{\max}/c_2)/(\alpha^2\varepsilon^2)$ is enough to guarantee $\mathcal{R}_B(\hat{\pi}_{\mathrm{IS}}) \leq \varepsilon$ with high probability.

2.4.1 Regularity assumptions

Both regularity assumptions (R2) and (R1) are quite mild and expected to hold in practice. The uniform connectivity assumption (R2) guarantees that it is possible to transit from any node in the infection set O to the final snapshot, i.e., $\rho(s \to O) > 0$ for all $s \in O$. If this assumption fails to hold for some nodes U, i.e., $\rho(s \to O) = 0$ for $s \in U$, we may restrict the estimation to the rest of nodes and the consistency holds for $\widehat{\ell}_{\mathrm{IS}}(s \to O)$ for $s \in O \setminus U$. When the network is unweighted, i.e., $A_{ij} \in \{0, 1\}$, one can take $c_2 = 1$ in (2.21) and (R2) just guarantees that O is connected. The general form of (R2) allows for arbitrarily small weights.

2.4.1.1 Connection to Szemerédi regularity

Condition (R1) is related to the notion of regularity introduced by Szemerédi. For some $\varepsilon > 0$, assume that O and O^c form an ε-regular partition in the sense of Szemerédi regularity lemma [KS]. Then, for any $U \subset O$ with $|U| \geq \varepsilon|O|$, we have

$$\mathrm{vol}(U, O^c) \geq \Big(\frac{\mathrm{vol}(O, O^c)}{|O|} - \varepsilon|O^c|\Big)|U|.$$

Let $d_{\text{ave}} = \text{vol}(O, O^c)/|O|$ and $q = d_{\text{ave}}/|O^c|$, the average degree of O in O^c and the the density of the cut, respectively. Then, taking $\varepsilon = q/2$, we have

$$\text{vol}(U, O^c) \geq \frac{1}{2}|U|d_{\text{ave}}$$

for all $|U| \geq (q/2)|O|$, which implies (2.20) with $c_1 = 1/2$ since $q \leq 1$. Szemerédi regularity generally holds for random graphs. In fact, this type of regularity is often used to argue quasi-randomness in any large deterministic network, which is one interpretation of Szemerédi lemma.

2.4.2 A note about the unbiased importance sampler

As shown in Lemma 1, $\widehat{Y}_{\text{unb}}$ is an unbiased estimator of the likelihood. Therefore, $-\log(\widehat{Y}_{\text{unb}})$ may be a more natural estimator of ℓ. It has a nice asymptotic behavior, including consistency for any infection size. Despite its appeal, it is more challenging to achieve a theoretical bound for the small sample variance of $-\log(\widehat{Y}_{\text{unb}})$. As a result, it is ambiguous whether the unbiased estimator outperforms $\widehat{\ell}_{\text{IS}}$. To fill this gap, we provide empirical comparison between two estimators in the simulations section. The comparison shows that $-\log(\widehat{Y}_{\text{unb}})$ performs inferior to our proposed estimator, $\widehat{\ell}_{\text{IS}}$, in recovering the source node.

2.5 Simulations

The proposed Importance Sampling (IS) has a reasonable runtime compared to previous Monte Carlo solutions and shows a superior performance in source identification. In this section, we make a comparison based on these two measures on real and synthetic networks.

We consider the rank loss to evaluate the performance. Each method in our simulation outputs a permutation, ranking the nodes according to their likelihood of being the source. We evaluate the methods based on the expected rank R, the expectation of the rank of the actual source among the list of candidates. We normalize to get a number in [0,1], with zero corresponding to perfect recovery, i.e., we use $(R-1)/n$. As a measure of performance, rank is more appropriate than "distance to source", esp. in networks with the small-world effect

(i.e., most of the network being within a small distance of any node.)

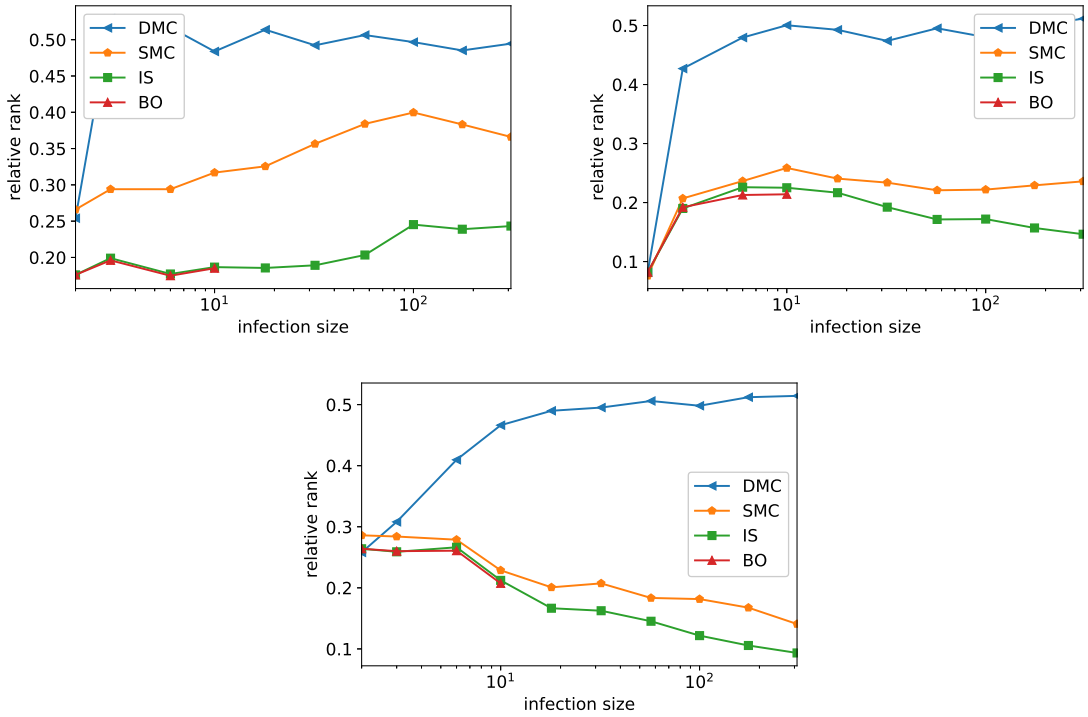


Figure 2.1: Plots of the expected relative rank versus the infection size for various networks: (left) DC-SBM, (right) Internet AS (bottom) US West Power Grid.

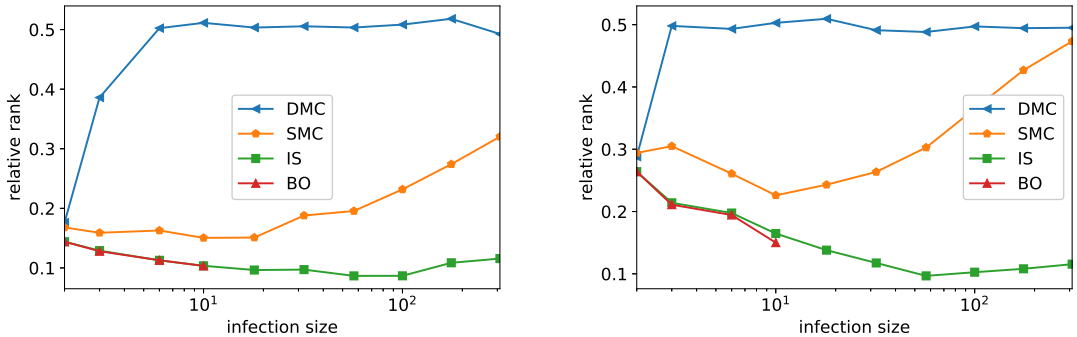


Figure 2.2: Plots of the expected relative rank versus the infection size (continued): (left) Wiki vote (right) UC64.

We consider a variety of real and simulated networks. Our selection includes an Internet Autonomous System [LKF05], US west-coast power grid [WS98], a Facebook-100

Table 2.1: Network statistics

Network	n	mean deg.	max deg.	clust. coeff.
Internet	10670	4.0	2312	0.01
Power	4941	3.0	19	0.1
Wiki vote	7066	29.0	1065	0.13
UC64	6810	46.0	660	0.19
DC-SBM	1962	66.0	897	0.3

networks [TKM11, TMP12], called UC64, and a Wikipedia voting network [LHK10]. In addition, we present our results on a synthetic network that is well-studied in the literature, namely, degree-correlated stochastic block model (DC-SBM) [KN11].

Table 2.1 summarizes the statistics on the largest connected component of these networks. For the DC-SBM network, we generate from a 3-community planted partition version, i.e., $\mathbb{E}[A_{ij}] = \theta_i\theta_j P_{ij}$ where $P_{ij} = 0.5$ if nodes i and j are in the same community and $P_{ij} = 0.02$ if they are in different communities. The degree parameters θ_i are generated from a rescaled Pareto distribution with $\alpha = 2$ and threshold $= 1$.

The results are illustrated in Figures 2.1 and 2.2. The methods we consider besides the Bayes Optimal solution (BO) and Importance Sampling (IS) are Direct Monte Carlo (DMC) and Soft Monte Carlo (SMC). We implement all the Monte Carlo methods using $N = 50$ draws. Each curve shows the performance of one method for different values of the infection size, in the range $2 \leq |O| \leq 300$. Each point is an average over 500 infection paths rooted at random sources (resulting in 500 random infection sets). To avoid an unreasonable computation time, we skip the BO for the infected sets of size greater than 10. The BO curve serves as the benchmark for the best achievable performance. Note that even the optimal solution needs to output a large set to catch the source, signifying the inherent difficulty of the problem.

The plots show that DMC is almost equivalent to random guessing for infection sizes

≥ 10. In contrast, SMC performs quite well for moderate infection sizes, but often starts to deteriorate quickly as the infection size grows.

Internet AS and power grid networks are the exceptions. These networks are less transitive (have lower clustering coefficient as shown in Table 2.1) and therefore the source recovery problem becomes simpler.

Importance Sampling is the global winner. Indeed, it virtually mimics the performance of the Bayes optimal estimator in all the plots. The results confirm Theorem 2 that IS accurately approximates the likelihood and therefore generates an efficient estimator of the source.

The behavior of the Bayes optimal estimator (or equivalently the ML estimator) was unknown for medium to large infection sizes. With the aid of the IS, we are capable of extrapolating the BO curve. One interesting finding about optimal source recovery in Figures 2.1 and 2.2 is that as the epidemics grows we capture more useful information about the actual source (relative to the size of snapshot). This counter-intuitive observation runs contrary to the perception that the source would conceal itself as the infection propagates throughout the network. If this finding turns out to be consistently true in highly transitive networks, it could revolutionize network forensics.

Figure 2.3 (left plot) illustrates the average runtimes for the infection size 100. The DMC is the fastest (due to the early stopping), followed by Importance sampling and SMC. In Figure 2.3 (right plot), we illustrate how IS, SMC, and DMC converge as a function of Monte Carlo sample size. We run 200 experiments on UC64 and fix the infection size at $|O| = 50$ and present the average rank. IS initiate with lower rank loss and converges faster. In the contrary, DMC never converges to the true likelihood, and SMC converges quite slowly. In Figure 2.3 (bottom plot), we plot the sampling variation of IS for a single epidemics of size 30 in UC64 network. Nodes are sorted in the order they are infected. The box plot over node i demonstrates the distribution of IS approximation for the log likelihood if i is the source. The sampling distribution is derived using 100 iterations of the IS algorithm. For this epidemics, The source node achieves the second highest IS value in median (after node 9).

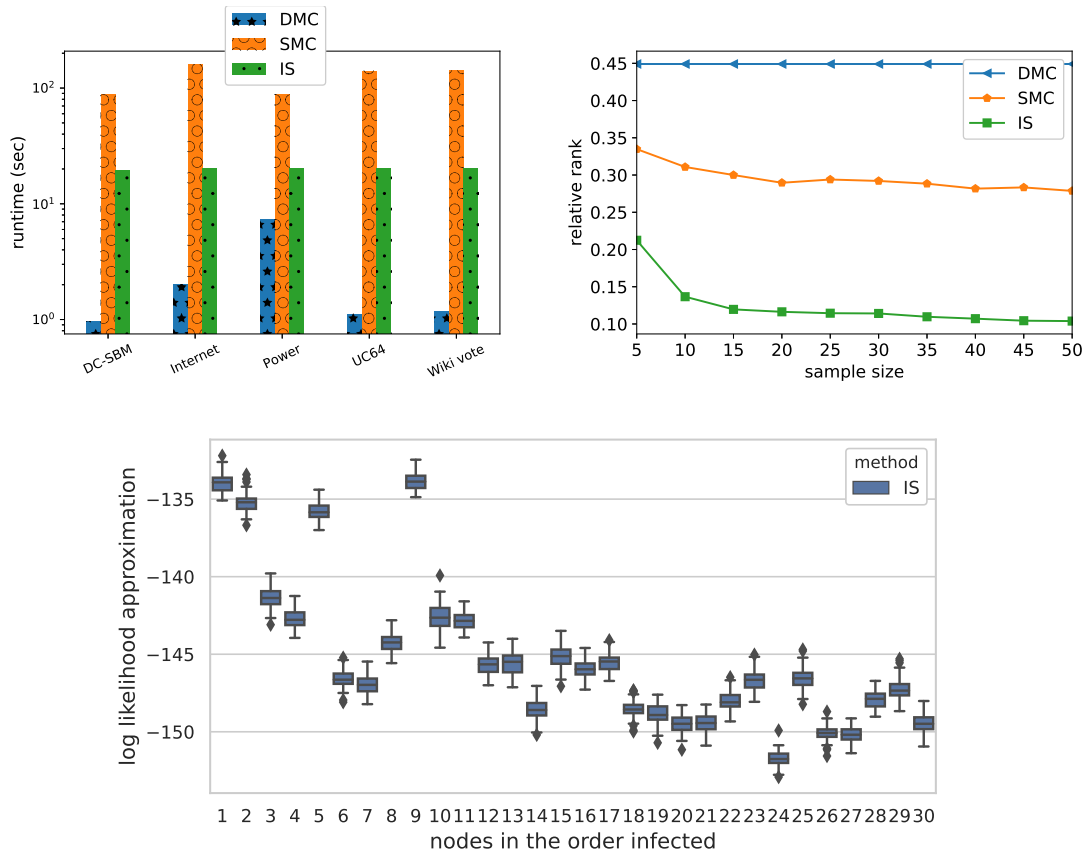


Figure 2.3: (Left) Average runtimes in seconds for infection size 100. (Right) Convergence rates for the expected relative rank as a function of the Monte Carlo sample size. (Bottom) Sampling variation of IS for a single epidemics of size 30.

Furthermore, the inter-quartile range is uniformly low, confirming the previous observation that IS converges with only 50 draws.

Based on these results, we advocate for the use of IS as the main tool for identifying sources of epidemics, regardless of the network topology or the nature of the epidemics (rumor propagation, disease contagion, etc.). It is fast and reliable, as supported by theory and simulations.

2.5.1 Comparison to the unbiased estimator

As described previously, $-\log(\widehat{Y}_{\text{unb}})$ is an alternative estimator to the negative log-likelihood of the source. We demonstrate empirically that it underperforms $\widehat{\ell}_{\text{IS}}$, our proposed estimator, in terms of rank loss.

Under similar empirical set-up explained earlier, we evaluate the unbiased IS. The results are illustrated in Figure 2.4. The infection size is limited to range below 100. As shown in the figure, IS globally generates lower rank loss than the unbiased IS.

In order to diagnose the reason behind the poor performance of the unbiased IS over IS, we compute its sampling variation for a single epidemics in UC64 network, identical to Figure 2.3. Figure 2.5 illustrates the comparison between two importance sampling methods. It shows the unbiased importance sampler generates noisy estimates of the log-likelihood as opposed to IS. We conclude that the bias-variance trade-off works in favor of IS in the source identification task.

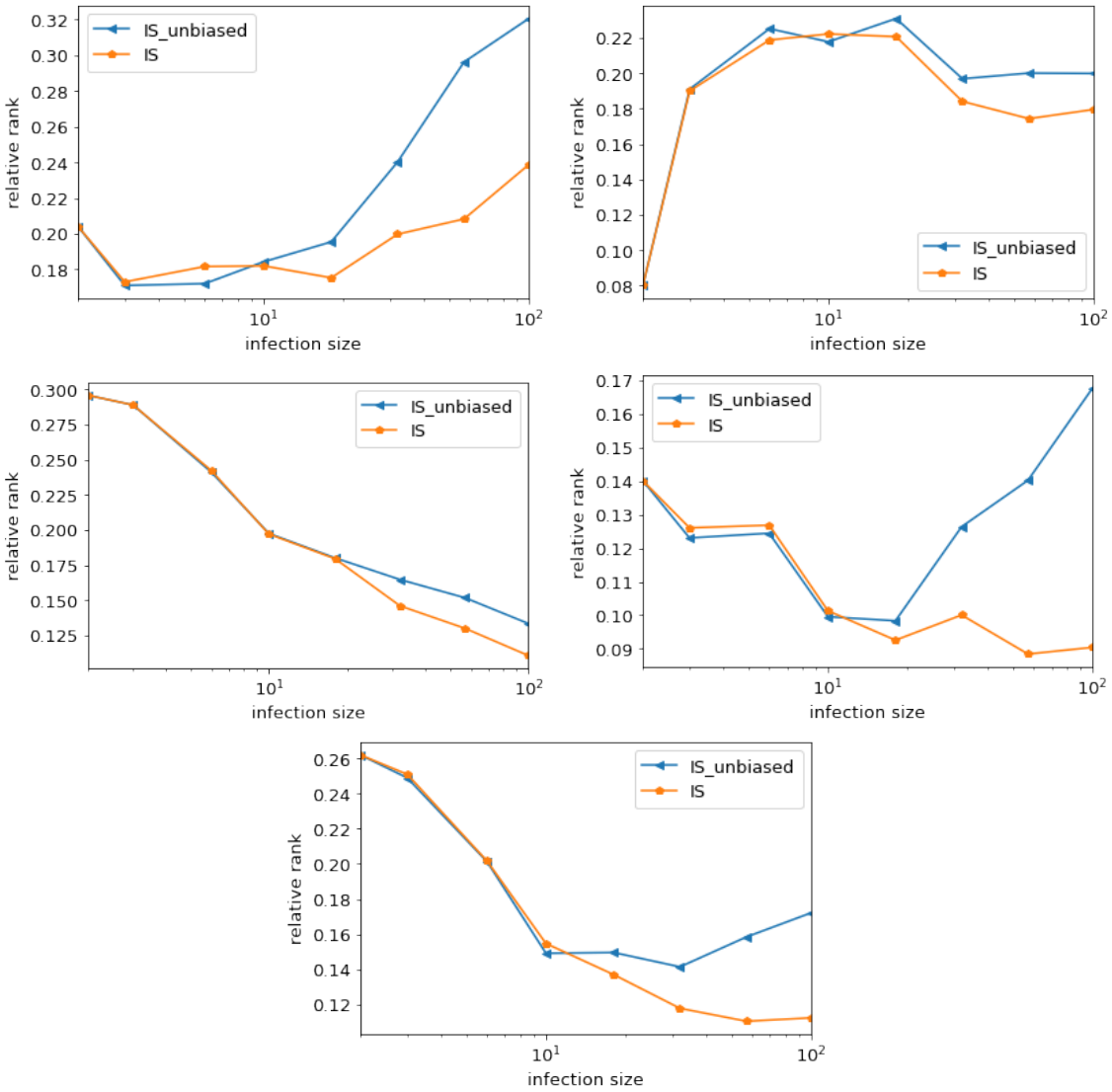


Figure 2.4: Evaluating the unbiased estimator. Plots of the expected relative rank versus the infection size for various networks: (top left) DC-SBM, (top right) Internet AS, (bottom left) US West Power Grid, (bottom right) Wiki vote (bottom center) UC64.

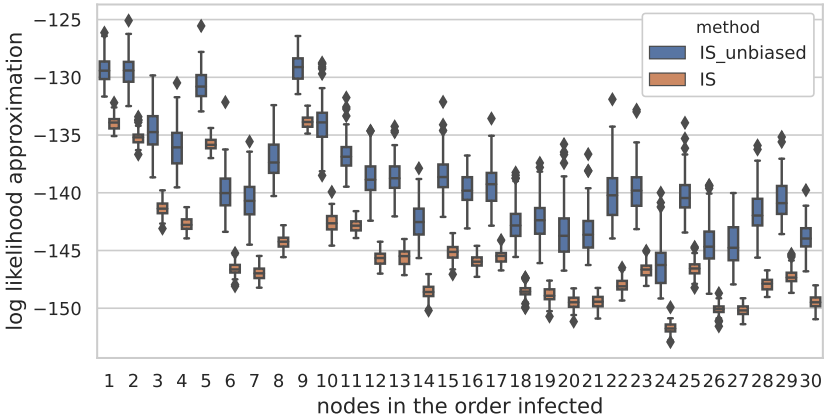


Figure 2.5: Sampling variation of IS and unbiased IS for a single epidemics of size 30.

Part II

Epidemic Network Reconstruction

CHAPTER 3

Monte Carlo Estimation

3.1 Introduction

The compartmental models have been utilized to explain how diseases spread across a population. The simple transition rules between compartments qualitatively describe the rise and fall of infection and the possibility of an outbreak. The model can provide numeric prediction of the compartment sizes as well, if accompanied with an estimation of the spread parameters. In well-mixed setup, where the nodes are assumed to be homogeneously connected, this estimation amounts to calibrating the final stage. In certain domains, such as influenza epidemics, numerous attempt led to accurate estimation of spread and recovery rates [AAM92].

In network epidemics, however, specially those with heterogeneous rates of spread, there are two challenges in estimating the model parameters. First, a high resolution data is hardly available. In absence of mega surveillance infrastructure (e.g. as in water distribution systems [OUS08]) or natural tools to trace back the propagation path (e.g. phylogenetic analysis of DNA evolutionary tree [DCB17]), we are deprived of the timestamp for individual infections and even the order of contamination in the network. Therefore, it is very common that the post-incidence observation is restricted to a snapshot of infected individuals. The most recent tools for the inference of epidemic rates on networks, however, require the full information about the propagation, including the individual links that transmitted the infection to the next victim. [KL18, GSD16]

Second challenge of inferring the epidemic parameters is the computational complexity of the inverse problem. Under various epidemic models and different observation granularities,

it has been shown that maximum likelihood estimation of the epidemic parameters is #P-complete [PV18, LTG10], unless the full transmission path is observed [KL18]. Filtering the epidemic output for static snapshots of the infection, introduces many hidden (masked) variables equal to the size of the problem. Such an inference problem with numerous hidden variables is inevitably treated with approximate inference methods such as EM, variational inference, or Monte Carlo methods [WJ08, BKM17, Sni02].

One of the imperative parameters of an epidemic is the underlying network through which the information is transmitted in varying spread rates. The task of discovering the spread rate in a pandemic may boil down to recovering the sparse topology of the epidemic network. An example is when one tries to recover the relevant social network from a microblogging news outbreak [AAR10]. Inferring the propagation rates in some other domains means to recover the link weights of the (potentially observed) epidemic network, such as to find the relevant cultro-geographic characteristics that facilitates the spread of a deadly disease across adjacent municipalities [LRB14].

Epidemic network reconstruction also reveals whether a viral adoption of an information, habit or disease is a result of contagious spread [KL17b], as opposed to the existence of an external "contaminating" source. This problem, whether a sparse network of contacts carries out the contagion or it is propagated independent of homophilic relations, has a close cousin. Given two possible epidemic networks which one is responsible for the viral spread [KL17b].

In this paper, we consider a version of the epidemic network reconstruction problem and propose a Monte Carlo estimation that addresses all challenges motioned above. In our setup, an infection propagates according to a continuous susceptible-infected dynamics with heterogeneous spread rates over a network of contacts. We assume, from a single realization of this epidemics, the set of infected nodes is observed at several points in time . Our Monte Carlo estimator then recovers the epidemic network. In case the network is known the method recovers its link weights, i.e. transmission rates.

This is the first time the epidemic network reconstruction is considered with such low level of observation granularity. Earlier contributions assume the luxury of observing the

entire timestamped infection sequence [GLK12] or the full transmission route, i.e. that who infected whom [KL18]. We should also note a line of literature that propose to recover the network from a set of connectivity constraints [AAR10]. However, those works ignore the dynamics of an epidemic and therefore stay away from the complexity of likelihood estimation in otherwise complicated settings.

We would like to enumerate the novelties in this chapter:

1. We propose a flexible compartmental model that allows inhomogeneous spread rates and external contamination.

2. We develop an importance sampling estimator that reconstruct the epidemic parameters from sparse cascade of snapshots, without knowledge of the infection times.

3. In a parametric setting, our estimator finds how much edge attributes matter for the flux of an infection.

4. Simulations demonstrate satisfactory estimation of the epidemic rates in a water distribution system.

The result of this paper can be applied to understand outbreaks and to develop strategies to prevent and contain hazardous epidemics. In many domains in public health, such as detecting constituents in water distribution system, multiple epidemics may occur on a single network. Hence, knowing the epidemic parameters guides us to plan for future outbreaks.

In what follows, we define the setup, including the continuous heterogeneous susceptible-infected epidemics with external contamination and argue for its generality. Then, the Bayes optimal solution to the network reconstruction problem is formalized. We present our importance sampling method afterwards. The paper concludes with a presentation of simulation results illustrating the performance of our method applied to water distribution system.

3.2 The Epidemic Model

Numerous attempts at explaining network epidemics has led to four main categories of information dynamics models: The Susceptible-infected (SI) model, the Information Cascade (IC) model, the Linear Threshold Model (LTM), and the Game Theoretic Model (GTM) [LWG17]. Among these, the SI dynamics and its variants have been used extensively in studying epidemics [KMS17] and have gained significant popularity. The model originates from population ODEs that describe the evolution of the infected population size [KM27]. It has been extended to Markov processes on networks that rule the spread of a disease (or social habit, information, etc.) from infected nodes to susceptible neighbors.

Some variations of the SI model include extra states in addition to being susceptible or infected. The most widely used are the Susceptible-Infected-Recovered (SIR) model, which allows for a recovery state after the infection, the Susceptible-Infected-Susceptible (SIS) model, where an infected node will recover without immunization, the SEIR (E stands for exposed), which adds a state of being infected but not infectious. Graph-coupled Hidden Markov Model (GCHMM) [DPH12] is another recent example, which follows a SIS (with external noise) model but takes the contagious infection as a hidden state. Some extensions introduce variations in the rate of the spread between neighboring nodes. The FSIR model [FHL15], for instance, assumes the rate of information spread is conversely proportional to the degree of a node, so that hubs become less prone to notice and retweet a message from a friend (due to limited attention). The ESIS model [WLJ15], on the other hand, introduces weights to the information spread according the emotional ties between friends and the sentiment of the message. See [LWG17, KMS17] for additional variations and extensions of the SI model.

In this paper, we assume that the epidemic is governed by a continuous heterogeneous contaminated susceptible-infected (CHCSI) dynamic. Let $G(V, E)$ be a connected, directed, and weighted graph defined by the set of vertices $V = [n]$ and the set of edges E, with adjacency matrix $A = (a_{ij})_{ij}$ representing the edge weights. Each node at continuous time t can be in an infected, $q_i(t) = 1$, or susceptible state, $q_i(t) = 0$. Infection propagates through two sources, via neighboring infected nodes and via an *external contamination*. The external

contamination uniformly exposes all susceptible nodes. As we model it, a susceptible node $i \in V$ is in risk of getting infected at the exponential rate $\lambda\,(b_i(t) + \beta)$, independent of other nodes. Here, $b_i(t) = \sum_{j \in V} a_{ji} q_j(t))$ denotes the (weighted) number of infected nodes neighbor to i at time t, $\beta > 0$ calibrates the strength of the external contamination and $\lambda > 0$ determines the rate of spread. Infection is an absorbing state for all nodes.

At time $t = 0$, all nodes are susceptible and the first infection is generated by the external contamination. As the infection grows, it finally covers the entire network. Let $\widetilde{\sigma} \in \mathrm{Sym}(V)$ denote the permutation of the nodes in the order they get infected. We refer to $\widetilde{\sigma}$ as the *infection path*. Here, $\mathrm{Sym}(V)$ refers to the set of permutations of V. The infection path does not reveal any information about the timing of infection. We write

$$\widetilde{\sigma} \sim \text{CHC-SIP}(G, \beta) \tag{3.1}$$

to denote the probability distribution for $\widetilde{\sigma}$. CHC-SIP stands for continuous (heterogeneous contaminated) susceptible-infected path. In the next section, we will characterize CHC-SIP and show that it is independent of λ.

We use $\widetilde{\sigma}_i$ to denote the i-th infected node and $\widetilde{\sigma}_{[i]}$ to denote the infection path truncated at its i-th element, with the convention that $\widetilde{\sigma}_{[0]} = \emptyset$. Also let $\widetilde{\sigma}_{[i,j]}$ refer to the subpath that starts from the $(i+1)$th and ends at the j-th element of $\widetilde{\sigma}$. In extreme cases, we set the convention that $\widetilde{\sigma}_{[0,i]} = \widetilde{\sigma}_{[i]}$ and $\widetilde{\sigma}_{[i,i]} = \emptyset$. For any sequence $V_1 \subset \cdots \subset V_m$, and $k_i = |V_i|,\ i \in [m]$, we use $\mathrm{Sym}((V_i)_{i=1}^m)$ to denote the permutations σ of V_m that conform to all V_i's, i.e. $\sigma_{[k_i]} \in \mathrm{Sym}(V_i),\ i \in [m]$.

Let $\widetilde{O}_i$ be the set of the first i infected nodes and call it the *infection set* or *snapshot* of size i. In other words,

$$\widetilde{O}_i = \{\widetilde{\sigma}_j \,:\, 1 \le j \le i\} \tag{3.2}$$

The event that $\{\widetilde{O}_i = O\}$ implies that all $j \in O$ are infected prior to any $j \in V \backslash O$. The infection set hides the information about the order in which nodes are infected. We define a *cascade* of snapshots as a series of snapshots $(\widetilde{O}_{i_1}, \ldots, \widetilde{O}_{i_m})$ where $1 \le i_1 < \cdots < i_m \le n$. Obviously, for a cascade, we have $\widetilde{O}_{i_1} \subset \cdots \subset \widetilde{O}_{i_m}$. The event that $(O_1, \ldots, O_m)$ is a cascade

of infection sets implies that the nodes in O_i are infected before the nodes in O_{i+1} for all $i \in [m-1]$ and the nodes in O_m are infected prior to the rest of the nodes in V.

In a network which is not instantaneously monitored, what we can observe from an epidemic is a single (or a cascade of) snapshot(s). In the next section, we will discuss the maximum likelihood approach to inferring parameters of an epidemic and the underlying network G from a single or a cascade of snapshots.

Before proceeding to the details of the methods, let us pause for a moment to discuss the generality of the continuous heterogeneous contaminated susceptible infected dynamic. CHCSI is a flexible extension of the SI dynamic that captures the variety in the rate at which the information/disease is transmitted from one neighbor to another. Examples of various factors that cause this distinction across the edges of the network are: the distances between cities, the length of water pipes or electric wires, the strength of emotional ties between friends, the limited attention span due to the abundance of friends, different levels of immunity to an infection among citizens, and so on. In this respect CH-SI generalizes the FSI and ESI, i.e., the FSIR and ESIR without the recovery state, and allows one to incorporate node and edge attributes in characterizing the epidemics. We will discuss, in the next section, how these attributes are linked to the speed of the information transmission.

We abide by the continuous nature of the original SI dynamic in this paper. Many prior works mentioned in the introduction adopt a discrete version of SI. In this respect, we solve the epidemic inference problem in a more general and realistic setup. As we will see in the next section, arguing in infinitesimal time steps, the solution can be relaxed to one degree of freedom, since the time parameter, i.e., when we happen to observe the snapshot, will disappear from the equations and is no longer required to be known a priori as in [GLK12] (or otherwise estimated as in [LMO14]).

We have incorporated external contamination to explain hidden nodes in the contact network (unobserved influencers). As a consequence, CHCSI is more robust to missing data than a simple SI dynamics. The independent external contaminator generates distinct patterns in the infection spread compared to a "real" epidemics. For example, the ordering

of infected nodes may not follow that of a epidemics and the infection set may become disconnected at different points of the epidemics. It worth noting that the building blocks of CHCSI has been proposed in early papers ([KL17b] for contamination and [JS08] for continuous heterogenous SI), but no other paper that we know of has assembled them into one model.

A drawback of CHCSI is the absence of recovery state. In many real situations, it is natural to think of individuals as developing immunization to a disease, abandoning a habit, or removing a tweet from their account. While it is beneficial to add the recovery state, it is conceivable that in a short period after the outbreak, no one develops immunization or abandons propagating the information. We hope in future we can extend the methods developed here to an epidemics with recovery state.

3.3 Epidemic Inference from Cascade of Snapshots

We now formulate the network recovery problem that can be efficiently solved by our sampling approach. Suppose that an epidemic outbreak starts in network G, and propagates according to a CHCSI dynamic, with a rate controlled by β, λ, and A, the weighted (asymmetric) adjacency of G. Assume every node tuple (i, j) is attributed with a vector $\tau_{ij} \in \mathbb{R}^{r_e}$. These attributes are observable but the adjacency is only known to follow:

$$a_{ij} = \varphi_\theta(\tau_{ij}), \quad \theta \in \Theta \tag{3.3}$$

for some $\theta = \theta^*$. We use $G = G(\theta)$ to denote a graph with its edges following (3.3). At different points in time, we observe the full snapshots of the infected nodes to be $O_1 \subset \cdots \subset O_m \subset V$. The task is to infer θ^* and the ensuing network $G(\theta)$.

We begin by defining the probability of the event that $\mathbf{O} = (O_i)_{i=0}^m$ is a cascade of snapshots of the epidemic, knowing that $\emptyset = O_0 \subsetneq O_1 \subsetneq \cdots \subsetneq O_m \subset V$. Suppose $k_i = |O_i|$, $0 \leq i \leq m$, and consider a "candidate θ" in (3.3). Let us write $\mathbb{P}_{\theta,\beta}(\cdot)$ to denote the distribution (3.1) of

a β-contaminated infection path in $G = G(\theta)$. Then,

$$\begin{aligned}\rho_{\to\mathbf{O}}(\theta,\beta) &:= \mathbb{P}_{\theta,\beta}(\widetilde{O}_{k_i} = O_i,\ 0 \le i \le m)\\ &= \sum_{\sigma\in\mathrm{Sym}(\mathbf{O})} \mathbb{P}_{\theta,\beta}(\widetilde{\sigma}_{[k_m]} = \sigma)\end{aligned} \tag{3.4}$$

is the probability of transition from O_0 to O_1, from O_1 to O_2, and etc. The sum on the RHS potentially contains $\Theta(\prod_{i=0}^{m-1} |\mathrm{Sym}(O_{i+1} \setminus O_i)|)$ nonzero terms. Even if G_{O_m} is a tree, the complexity of computing this formula can be non-polynomial in $k_m - m$. The difficulty of solving epidemic network recovery is due to the complexity of evaluating this sum for a general network topology.

Let us introduce some notation: Let $\mathrm{vol}_G(U,W) = \sum_{u\in U, w\in W} a_{uw}$ be the the weight of the cut between nodes U and W in the network G. We also write $\partial\mathrm{vol}_G(U,W) = \mathrm{vol}_G(U, W\setminus U)$. For singleton sets $\{i\}$, we often drop the braces and write i.

As was previously shown [CZC15b, KA19, KL17b], the odds of a certain infection path governed by the SI dynamic has a time and rate-free form and can be computed in linear time. The same applies to CHC-SIP pmf. For a $\sigma \in \mathrm{Sym}(\mathbf{O})$,

$$\begin{aligned}\mathbb{P}_{\theta,\beta}(\widetilde{\sigma}_{[k_m]} = \sigma) &= \prod_{i=0}^{k_m-1} \mathbb{P}_{\theta,\beta}(\widetilde{\sigma}_{i+1} = \sigma_{i+1} \mid \widetilde{\sigma}_{[i]} = \sigma_{[i]})\\ &= \prod_{i=0}^{k_m-1} \frac{\partial\mathrm{vol}_{G(\theta)}(\sigma_{[i]}, \sigma_{[i+1]}) + \beta}{\partial\mathrm{vol}_{G(\theta)}(\sigma_{[i]}, [n]) + (n-i)\beta}\end{aligned} \tag{3.5}$$

Note that $\partial\mathrm{vol}(\sigma_{[i]}, \sigma_{[i+1]}) = \mathrm{vol}(\sigma_{[i]}, \sigma_{i+1})$. This equation follows easily using the Markov property of the dynamics and a classic property regarding the minimum of independent exponential random variables [Ros14]. Note that (3.5) does not depend on either the rate λ or the time of observation.

With the cascade $\mathbf{O} = (O_i)_{i=1}^m$ as the observation, the maximum likelihood estimator for the θ^* follows:

$$\widehat{(\theta,\beta)}_{\mathrm{ML}}(\mathbf{O}) := \operatorname*{argmax}_{\theta,\beta} \log \mathbb{P}_{\theta,\beta}(\widetilde{O}_{k_i} = O_i,\ 1 \le i \le m). \tag{3.6}$$

In the next section, we discuss Monte Carlo solutions for approximating the likelihood of a cascade of complete snapshots.

3.4 Monte Carlo Estimation

We now derive Monte Carlo estimates for $\mathbb{P}_{\theta,\beta}(\widetilde{O}_{k_i} = O_i,\ 1 \le i \le m)$ for any cascade of snapshots $\mathbf{O} = (O_i)_{i=1}^m$. We describe our main contribution in this paper, the Importance Sampling (IS), and will prove a guarantee for its fast convergence to the likelihood.

3.4.1 Importance sampling

Since $\rho_{\to \mathbf{O}}$ is quite small for large k, directly sampling from the target distribution generates zero term in Naive Monte Carlo sampling in the majority of iterations. That is, is it very unlikely that a path generated from the dynamic on the whole network conforms entirely to $\mathbf{O}$. The idea which falls under the umbrella of importance sampling [Gew89] is to tweak the proposal distribution to avoid these zero contributions. The scheme we propose is to constrain the epidemic spread to $G_{O_1}, \ldots, G_{O_m}$ and draw infection paths accordingly. These paths are by design guaranteed to always hit $\mathbf{O}$, as opposed to those generated by the Naive Monte Carlo.

More specifically, let $\theta_0 \in \Theta$ be another parametrization for the adjacency in (3.3) and $\beta_0 > 0$ be an arbitrary contamination parameter. We draw samples $\{\widetilde{\sigma}^t\}$ from the β_0-contaminated CHC-SIP dynamic restricted to the subgraphs of $G(\theta_0)$ on O_i's, denoted as $G_{O_i}(\theta_0),\ i \in [m]$. That is, we draw

$$\widetilde{\sigma}^t \sim \text{CHC-SIP}((G_{O_i}(\theta_0))_{i=0}^m, \beta),\ t = 1, \ldots, N, \tag{3.7}$$

independently. The distribution in (3.7) dictates that $\widetilde{\sigma}^t_{[k_i,k_{i+1}]}$ grows according to CHCSIP in $G_{O_{i+1}}$. We consider the estimator

$$\begin{aligned} \widehat{Y}_{\text{unb}} &= \frac{1}{N}\sum_{t=1}^{N} Y^{\theta,\beta}_{\theta_0,\beta_0}(\widetilde{\sigma}^t), \\ Y^{\theta,\beta}_{\theta_0,\beta_0}(\widetilde{\sigma}^t) &:= \prod_{i=0}^{m-1}\prod_{j=k_i}^{k_{i+1}-1} \frac{\partial \text{vol}_{G(\theta_0)}(\widetilde{\sigma}^t_{[j]}, O_{i+1}) + (k_{i+1}-j)\beta_0}{\partial \text{vol}_{G(\theta)}(\widetilde{\sigma}^t_{[j]}, [n]) + (n-j)\beta} \cdot \frac{\partial \text{vol}_{G(\theta)}(\widetilde{\sigma}_{[j]}, \widetilde{\sigma}_{[j+1]}) + \beta}{\partial \text{vol}_{G(\theta_0)}(\widetilde{\sigma}_{[j]}, \widetilde{\sigma}_{[j+1]}) + \beta_0}. \end{aligned} \tag{3.8}$$

Let $\widetilde{\sigma} \sim \text{CHC-SIP}((G_{O_i}(\theta_0))_{i=0}^m, \beta)$ be a generic sample from the restricted dynamic. We denote the distribution of $\widetilde{\sigma}$ as $\mathbb{P}_{\beta_0}^{G_{\mathbf{O}}(\theta_0)}$ to emphasize that it is a β_0-contaminated dynamic

restricted to $(G_{O_i}(\theta_0))_{i=0}^m$. We have

$$\mathbb{P}_{\beta_0}^{G_{\mathbf{O}}(\theta_0)}(\widetilde{\sigma} = \sigma) = \prod_{i=0}^{m-1} \prod_{j=k_i}^{k_{i+1}-1} \frac{\partial \mathrm{vol}_{G(\theta_0)}(\sigma_{[j]}, \sigma_{[j+1]}) + \beta_0}{\partial \mathrm{vol}_{G(\theta_0)}(\sigma_{[j]}, O_{i+1}) + (k_{i+1} - j)\beta_0}$$

Then,

$$\begin{aligned}
\mathbb{E}[\widehat{Y}_{\mathrm{unb}}] &= \mathbb{E}[Y_{\theta_0,\beta_0}^{\theta,\beta}(\widetilde{\sigma})] \\
&= \sum_{\sigma \in \mathrm{Sym}(\mathbf{O})} \mathbb{P}_{\beta_0}^{G_{\mathbf{O}}(\theta_0)}(\widetilde{\sigma} = \sigma) Y_{\theta_0,\beta_0}^{\theta,\beta}(\sigma) \\
&= \sum_{\sigma \in \mathrm{Sym}(\mathbf{O})} \prod_{i=0}^{m-1} \prod_{j=k_i}^{k_{i+1}-1} \Big[\frac{\partial \mathrm{vol}_{G(\theta_0)}(\sigma_{[j]}, \sigma_{[j+1]}) + \beta_0}{\partial \mathrm{vol}_{G(\theta_0)}(\sigma_{[j]}, O_{i+1}) + (k_{i+1} - j)\beta_0} \\
&\qquad \cdot \frac{\partial \mathrm{vol}_{G(\theta_0)}(\widetilde{\sigma}^t_{[j]}, O_{i+1}) + (k_{i+1} - j)\beta_0}{\partial \mathrm{vol}_{G(\theta)}(\widetilde{\sigma}^t_{[j]}, [n]) + (n - j)\beta} \\
&\qquad \cdot \frac{\partial \mathrm{vol}_{G(\theta)}(\sigma_{[j]}, \sigma_{[j+1]}) + \beta}{\partial \mathrm{vol}_{G(\theta_0)}(\sigma_{[j]}, \sigma_{[j+1]}) + \beta_0} \Big] \\
&= \sum_{\sigma \in \mathrm{Sym}(\mathbf{O})} \prod_{i=0}^{m-1} \prod_{j=k_i}^{k_{i+1}-1} \frac{\partial \mathrm{vol}_{G(\theta)}(\sigma_{[j]}, \sigma_{[j+1]}) + \beta}{\partial \mathrm{vol}_{G(\theta)}(\widetilde{\sigma}^t_{[j]}, [n]) + (n - j)\beta} \\
&= \sum_{\sigma \in \mathrm{Sym}(\mathbf{O})} \mathbb{P}_{\theta,\beta}(\widetilde{\sigma} = \sigma) = \rho_{\to \mathbf{O}}(\theta, \beta).
\end{aligned}$$

showing that $\widehat{Y}_{\mathrm{unb}}$ is an unbiased estimator of $\rho_{\to \mathbf{O}}(\theta, \beta)$.

Our proposed estimator is a modification of $\widehat{Y}_{\mathrm{unb}}$, namely,

$$\widehat{\ell}_{\mathrm{IS}}(\theta, \beta; \theta_0, \beta_0) := \frac{1}{N} \sum_{i=1}^{N} - \log Y_{\theta_0,\beta_0}^{\theta,\beta}(\widetilde{\sigma}^t), \tag{3.9}$$

which is an estimator of the negative log-likelihood $\ell_O(\theta) := -\log \rho_{\to \mathbf{O}}(\theta)$. That is, instead of estimating the probabilities directly, we estimate their log. This estimator is no longer unbiased. However, it has the advantage of having a much lower variance, hence requires much smaller Monte Carlo sample size (N) to achieve a given accuracy.

3.4.2 Approximate inference

A natural next step to estimate epidemic parameters, θ^*, β^*, is to minimize the importance sampling approximation for the negative log-likelihood defined in (3.9). Starting from initial

candidates θ_0, β_0,

$$\widehat{(\theta, \beta)}_{\mathrm{IS}} := \operatorname*{argmin}_{\theta,\beta} \widehat{\ell}_{\mathrm{IS}}(\theta, \beta; \theta_0, \beta_0) \tag{3.10}$$

is our proposed approximate MLE solution. Its accuracy depends on the Monte Carlo sample size and how close were our initial parameters to the final estimate. To make $\widehat{(\theta, \beta)}_{\mathrm{IS}}$ a better approximation of $\widehat{(\theta, \beta)}_{\mathrm{ML}}$, one could either increase N or iteratively update θ_0, β_0 to $\widehat{(\theta, \beta)}_{\mathrm{IS}}$, resample and solve (3.10). We leave it to future works to investigate the efficacy of these approaches.

Problem (3.10) is not a convex optimization. Upon a convenient choice for φ_θ, the objective becomes a difference of convex functions as shown below:

$$\begin{aligned}\widehat{(\theta, \beta)}_{\mathrm{IS}} = \operatorname*{argmin}_{\theta,\beta} \sum_{i=0}^{m-1} \sum_{j=k_i}^{k_{i+1}-1} & \log\Big(\partial \mathrm{vol}_{G(\theta)}(\widetilde{\sigma}^t_{[j]}, [n]) + (n-j)\beta\Big) \\ & - \log\Big(\partial \mathrm{vol}_{G(\theta)}(\widetilde{\sigma}_{[j]}, \widetilde{\sigma}_{[j+1]}) + \beta\Big)\end{aligned} \tag{3.11}$$

One candidate for the parametric family is

$$a_{ij} = \varphi_\theta(\tau_{ij}) = \exp(\theta^T \tau_{ij})$$

If we also replace β with e^β, (3.11) becomes a DC problem in (θ, β).

3.5 Simulations

In the last section we have developed an estimator (3.10) for transmission weights in an epidemic network with edge attributes. In this section, we assess its performance on real and synthetic networks. We use quadratic loss, i.e. bias squared plus variance, to evaluate our estimator.

We consider the water distribution network used as an example in EPANET tutorial [Ros00]. The subgraph we use consists of 92 nodes and 113 edges, illustrated in Figure 3.1. We consider the length of pipes as the only edge attribute determining the rate of spread between end points. Once normalized to fall within $[1, 2]$, let τ_{ij} denote the (i, j) edge attribute. The transmission (spread) rate is defined to be $a_{ij} = e^{\theta \tau_{ij}}$. We take $\theta = 8$ to

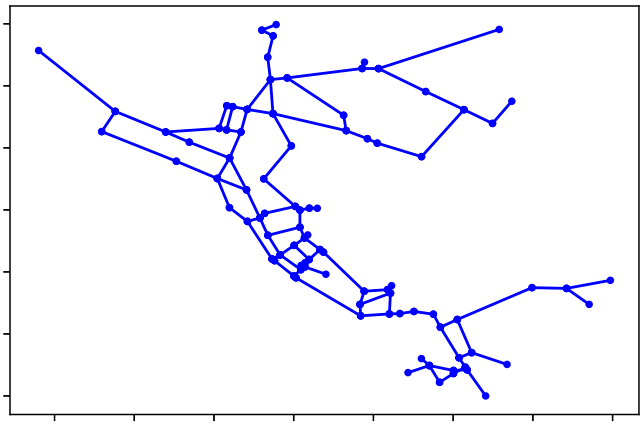


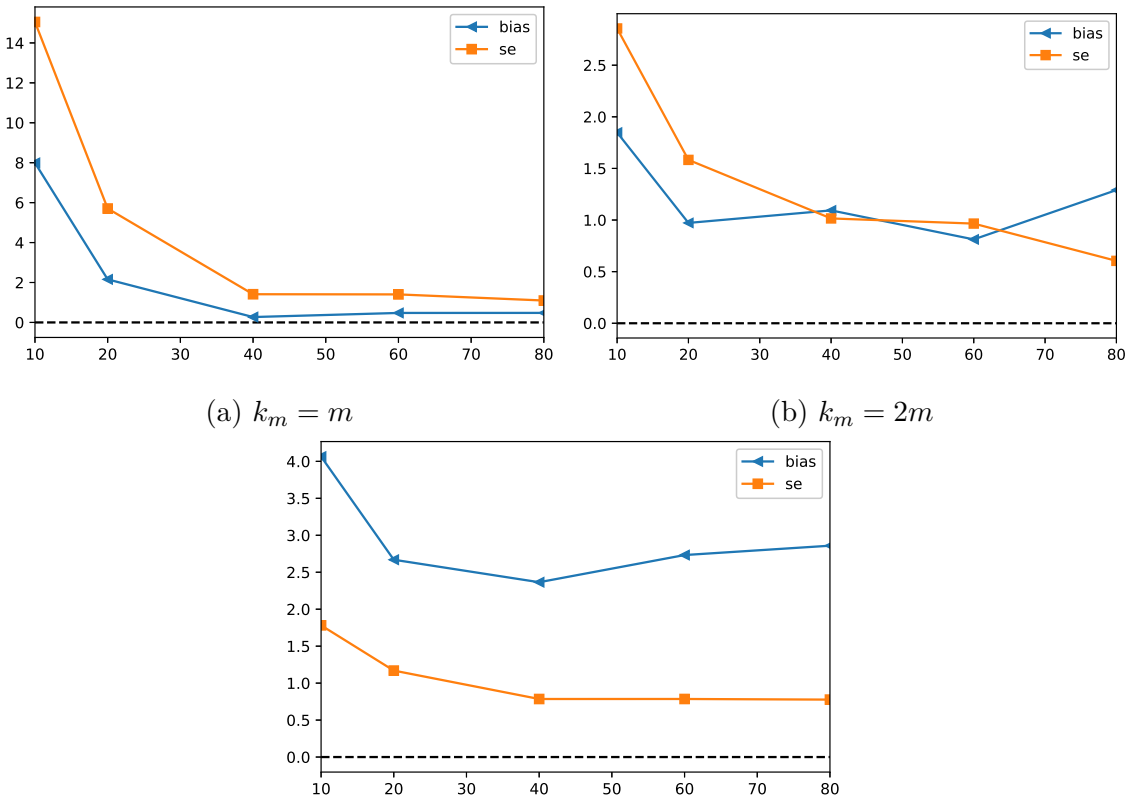
Figure 3.1: EPANET [Ros00] water distribution network

introduce sufficient amount of variation in likelihood necessary for accurate estimation. The contamination parameter is $\beta = 1$.

We test the performance of $\widehat{\theta}_{\text{IS}}$ for different infection sizes $10 \leq k_m \leq 80$ and multiple cascade sizes, m (the number of snapshots observed). We set $m = k_m/c$, for $c = 1, 2, 5$. For each combination of infection and cascade sizes we generate 30 random epidemics to compute the sampling bias and the standard error of the importance sampler estimator. The estimator is set to take $k_m - m$ Monte Carlo samples.

Figures 3.2a, 3.2b, 3.2c illustrate the performance of $\widehat{\theta}_{\text{IS}}$ on EPANET WDS. Each one shows the bias and standard error for a range of infection sizes. In Figure 3.2a, the cascade size is equal to infection size for each point on the curves. In other terms, the entire infection path is observed. In Figures 3.2b and 3.2c the infection set is observed less often, particularly after every 2 or 5 infections, respectively.

As expected before, the method is not asymptotically consistent, meaning the quadratic loss does not converge to zero as $k_m \to 0$. However, the estimator becomes less volatile and the bias remains in a satisfactory limit. Unfortunately, $\widehat{\beta}_{\text{IS}}$ is not performing well, hence we have decided not to display its results.



(a) $k_m = m$

(b) $k_m = 2m$

(c) $k_m = 5m$

Figure 3.2: Plots of the bias and standard error on Epanet WDS

Part III

Appendices

CHAPTER 4

Variational Source Identification

4.1 Multi-source Extension

The inference problem discussed in Section 1.2.2 immediately extends to the multi-source situations. Consider the case were more than one independent source, denoted by $\boldsymbol{I}_*$, initiate the infection dynamics. Due to the Markovian nature of the dynamics, the infection path that leads to some set I does not influence the value of $\rho_{I\to O}$. Hence, Proposition 1 also describes the likelihood of the transition from the source set $\boldsymbol{I}_*$ to a snapshot O.

If we know that there are s original sources, e.g. $|\boldsymbol{I}_*| = s$, with a uniform prior on the patient zeros, the Bayesian solution would be characterized by the optimization

$$I^*_{\text{MAP}} = \underset{I\subset O,\,|I|=s}{\operatorname{argmax}}\ \rho_{I\to O} \tag{4.1}$$

To compute this MAP estimate, we can still use the DP solution in Proposition 1, but we do not need to compute $\rho_{I\to O}$ for $|I| < s$. Thus, the multi-source problem is in a sense "easier", especially when $s \approx |O|$, since one can terminate the recursion earlier (i.e., the case $s = 1$ is the hardest).

4.2 Proofs

4.2.1 Proof of Proposition 1

Let us first recall a known fact about the exponential distribution:

Lemma 2. *Let $T_i \sim \mathrm{Exp}(\beta_i)$ be a collection of independent exponential variables. Then,*

$$\mathbb{P}\Big(T_i < \min_{j\neq i} T_j\Big) = \frac{\beta_i}{\sum_j \beta_j}.$$

For a proof of Lemma 2, see [Ros14]. The forward programming (1.2) is an application of the law of total probability in the following sense: The event that nodes in $O \setminus I$ are infected before any other node in I^c splits into sub-events that each node in $O \setminus I$ is infected before those in O^c and we have

$$\rho_{I\to O} = \sum_{j\in O\setminus I} \rho_{I\to I\cup j} \cdot \rho_{I\cup j\to O}$$

where we have also used the Markov property of SI dynamics to split the probabilities on the RHS into the products. The ratio in (1.2) corresponds to the transition probability from I to $I \cup j$, that is $\rho_{I\to I\cup j}$. Indeed, given that I is infected, we run exponential clocks $T_j \sim \mathrm{Exp}(\beta\,\mathrm{vol}(I,j))$ and the first to expire determines the next infected node. By Fact 2, this happens for any node $j \in I^c$ with probability probability $\propto_j \beta\,\mathrm{vol}(I,j)$. Thus,

$$\rho_{I\to I\cup j} = \frac{\beta\,\mathrm{vol}(I,j)}{\sum_{j'}\beta\,\mathrm{vol}(I,j')} = \frac{\mathrm{vol}(I,j)}{\mathrm{vol}(I,I^c)}.$$

This proves the forward programming. The backward programming, on the other hand, connects $\rho_{I\to O}$ to $\rho_{I\to O\setminus j}$ and is proved similarly. Basically, the event of visiting O can be divided into sub-events based on the last node in O that is infected.

4.2.2 Proof of Proposition 2

We prove the following alternative expressions for $S = (S_{jj'})^{|O|\times|O|}$ and $\boldsymbol{z} = (z_j)^{|O|}$,

$$S_{jj'} := \begin{cases} d^{in}_{O\setminus j'}(j)d^{in}_{O\setminus j}(j') + \sum_{i\in O} A_{ij}A_{ij'} & j \neq j' \\ 2\Big[d^{in}_{O}(j)^2 + \sum_{i\in O} A_{ij}^2\Big] & j = j' \end{cases}$$

$$\begin{aligned} z_j := &\Big[\mathrm{vol}(O_{\setminus j}) + 2\,\mathrm{vol}(O\setminus j, (O\setminus j)^c)\Big]\, d^{in}_O(j) \\ &+ \sum_{i\in O}(d^{out}_{O\setminus j}(i) - d^{in}_{O\setminus j}(i))A_{ij} + 2\sum_{i\in O} d^{out}_{(O\setminus j)^c}(i)\,A_{ij}. \end{aligned}$$

Here, $d_O^{out}(i) := \sum_{j\in O} A_{ij}$ is the out-degree of node i in O, $d_O^{in}(i) := \sum_{j\in O} A_{ji}$ is the in-degree of node i in O, and $\mathrm{vol}^{(2)}(i,j) := \sum_{r\in O} A_{ir}A_{rj}$ is the number of paths of length 2 between nodes i and j that pass through O. It is not hard to verify that these expressions are equivalent to the matrix form presented in (2).

Recall that $\mathrm{vol}(I, I^c) = \sum_{i,k} A_{ik}1\{i \in I, k \notin I\}$ and similarity $\mathrm{vol}(I, j) = \sum_r A_{rj}1\{r \in I\}$. Here, the indices, i, k and r run over all nodes in the network, i.e. $i, k, r \in [n]$. We have

$$\begin{aligned}
(Q^T\boldsymbol{r})_j &= \sum_{I\subset O} 1\{j \notin I\}\ \mathrm{vol}(I,j)\cdot\mathrm{vol}(I,I^c) \\
&= \sum_{I\subset O\setminus\{j\}} \mathrm{vol}(I,I^c)\cdot\mathrm{vol}(I,j) \\
&= \sum_{I\subset O\setminus\{j\}} \sum_{i,k,r} A_{ik}A_{rj}\,1\{i\in I,\ k\notin I,\ r\in I\} \\
&= \sum_{i,k,r} A_{ik}A_{rj}\gamma_{ikr}
\end{aligned}$$

where the last equality follows by interchanging the order of summations and defining

$$\gamma_{ikr} := \sum_{I\subset O\setminus\{j\}} 1\{i\in I,\ k\notin I,\ r\in I\}$$

If i or r do not belong to $O\setminus\{j\}$, or $k \in \{i, r\}$, then $\gamma_{ikr} = 0$. Thus, it what follows assume that $i, r \in O_{\setminus j} := O\setminus\{j\}$ and $k \notin \{i, r\}$. Then,

$$\gamma_{ikr} = \begin{cases} 2^{|O|-4} & i\neq r,\ k\in O_{\setminus j} \\ 2^{|O|-3} & i = r,\ k\in O_{\setminus j} \\ 2^{|O|-3} & i\neq r,\ k\notin O_{\setminus j} \\ 2^{|O|-2} & i = r,\ k\notin O_{\setminus j} \end{cases}$$

To see the second equality, note that we are counting subsets of the set $O\setminus\{j\}$ (of cardinality $|O|-1$) that contain or exclude certain elements. For example, when k, i, r are pairwise distinct, and $k \in O\setminus\{j\}$, looking at the binary representation of I, we have two ones in the positions i and r and a zero in position k, and the rest of $|O|-1-3$ positions are free to be zero or one.

In what follows, i and r range over $O \setminus \{j\}$ (otherwise $\gamma_{ikr} = 0$). Also, condition $k \notin \{i, r\}$ can be replaced with $k \neq r$, since the $k \neq i$ is implicitly enforced by $A_{ik} = 0$ if $k = i$ (no self-loops). We have

$$\begin{aligned}(Q^T \boldsymbol{r})_j &= \sum_{i,r} \sum_{k \neq r} A_{ik} A_{rj} \Big[2^{|O|-4} (1 + 1\{i = r\}) 1\{k \in O_{\setminus j}\} \\ &\quad + 2^{|O|-3} (1 + 1\{i = r\}) 1\{k \notin O_{\setminus j}\} \Big] \\ &= 2^{|O|-4} \sum_{i,r} d^{out}_{O \setminus \{j,r\}}(i) A_{rj} (1 + 1\{i = r\}) \\ &\quad + 2^{|O|-3} \sum_{i,r} d^{out}_{(O \setminus j)^c}(i) A_{rj} (1 + 1\{i = r\})\end{aligned}$$

where in the second term, we used the fact that if $k \notin O_{\setminus j}$ then we automatically have $k \neq r$ since r ranges over $O_{\setminus j}$. We have

$$\begin{aligned}\sum_r d^{out}_{O \setminus \{j,r\}}(i) A_{rj} &= \sum_r (d^{out}_{O_{\setminus j}}(i) - A_{ir}) A_{rj} \\ &= d^{out}_{O_{\setminus j}}(i) d^{in}_{O_{\setminus j}}(j) - \mathrm{vol}^{(2)}_{O_{\setminus j}}(i, j)\end{aligned}$$

where $\mathrm{vol}^{(2)}_{O_{\setminus j}}(i, j) := \sum_{r \in O_{\setminus j}} A_{ir} A_{rj}$ is the number of paths of length two between i and j in $O_{\setminus j}$. Note that $\mathrm{vol}^{(2)}_{O_{\setminus j}}(i, j) = \mathrm{vol}^{(2)}_O(i, j)$ and similarly $d_{O_{\setminus j}}(j) = d_O(j)$ since $A_{jj} = 0$. Thus,

$$\begin{aligned}\sum_{i,r} d^{out}_{O \setminus \{j,r\}}(i)\, A_{rj} \Big(1 + 1\{i = r\}\Big) &= \sum_i \Big[d^{out}_{O_{\setminus j}}(i) d^{in}_O(j) - \mathrm{vol}^{(2)}_O(i, j) + d^{out}_{O_{\setminus j}}(i) A_{ij} \Big] \\ &= \sum_i d^{out}_{O_{\setminus j}}(i) d^{in}_O(j) + (d^{out}_{O_{\setminus j}}(i) - d^{in}_{O_{\setminus j}}(i)) A_{ij} \\ &= \mathrm{vol}(O_{\setminus j}) d^{in}_O(j) + \sum_i (d^{out}_{O_{\setminus j}}(i) - d^{in}_{O_{\setminus j}}(i)) A_{ij}\end{aligned}$$

where $\mathrm{vol}(O_{\setminus j}) = \mathrm{vol}(O_{\setminus j}, O_{\setminus j})$ and the third equality follows since we have

$$\sum_{i \in A} \mathrm{vol}^{(2)}_A(i, j) = \sum_{i \in A} \sum_{r \in A} A_{ir} A_{rj} = \sum_{r \in A} d^{in}_A(r) A_{rj}$$

which was used with $A = O_{\setminus j}$. Similarly, we have

$$\begin{aligned}\sum_{i,r} d^{out}_{(O \setminus j)^c}(i) A_{rj} (1 + 1\{i = r\}) &= \sum_i d^{out}_{(O \setminus j)^c}(i) \Big(d^{in}_{O \setminus j}(j) + A_{ij} \Big) \\ &= \mathrm{vol}(O \setminus j, (O \setminus j)^c)\, d^{in}_O(j) \\ &\quad + \sum_i d^{out}_{(O \setminus j)^c}(i)\, A_{ij}\end{aligned}$$

It follows that

$$(Q^T \boldsymbol{r})_j = 2^{|O|-4}\Big[\operatorname{vol}(O_{\setminus j}) d_O^{in}(j) + \sum_i (d_{O_{\setminus j}}^{out}(i) - d_{O_{\setminus j}}^{in}(i)) A_{ij} \\ + 2\operatorname{vol}(O\setminus j, (O\setminus j)^c)\, d_O^{in}(j) + 2\sum_i d_{(O\setminus j)^c}^{out}(i)\, A_{ij} \Big].$$

Calculating Q^TQ Let us first take $j \neq j'$. Then, similar to the previous argument,

$$\begin{aligned}(Q^TQ)_{jj'} &= \sum_{I\subset O\setminus\{j,j'\}} \operatorname{vol}(I,j)\ \operatorname{vol}(I,j') \\ &= \sum_{I\subset O\setminus\{j,j'\}} \sum_{i,r} A_{ij}\, A_{rj'} 1\{i\in I,\ r\in I\} \\ &= \sum_{i,r} A_{ij}\, A_{rj'} \beta_{ir}\end{aligned}$$

where we have defined

$$\begin{aligned}\beta_{ir} &:= \sum_{I\subset O\setminus\{j,j'\}} 1\{i\in I,\ r\in I\} \\ &= 2^{|O|-4} 1\{i\neq r\} + 2^{|O|-3} 1\{i=r\} \\ &= 2^{|O|-4}\Big(1 + 1\{i=r\}\Big)\end{aligned}$$

assuming $i, r \in O\setminus\{j,j'\}$, otherwise $\beta_{ir} = 0$. Thus, restricting summations over indices $i, r \in O\setminus\{j,j'\}$

$$\begin{aligned}(Q^TQ)_{jj'} &= 2^{|O|-4}\Big[\sum_{i,r} A_{ij}\, A_{rj'} + \sum_i A_{ij}A_{ij'}\Big] \\ &= 2^{|O|-4}\Big[d_{O\setminus j'}^{in}(j) d_{O\setminus j}^{in}(j') + \sum_i A_{ij}A_{ij'}\Big].\end{aligned}$$

Now consider the case $j = j'$. Then,

$$\begin{aligned}(Q^TQ)_{jj} &= \sum_{I\subset O\setminus\{j\}} \operatorname{vol}(I,j)^2 \\ &= \sum_{I\subset O\setminus\{j\}} \sum_{i,r} A_{ij}\, A_{rj} 1\{i\in I,\ r\in I\} \\ &= \sum_{i,r} A_{ij}\, A_{rj}\, 2^{|O|-3}\Big(1 + 1\{i=r\}\Big),\end{aligned}$$

assuming $i, r \in O \setminus j$. It follows that

$$\begin{aligned}(Q^T Q)_{jj} &= 2^{|O|-3}\Big[\sum_{i,r} A_{ij}\, A_{rj} + \sum_{i} A_{ij}^2\Big] \\ &= 2^{|O|-3}\Big[d_O^{in}(j)^2 + \sum_{i} A_{ij}^2\Big].\end{aligned}$$

CHAPTER 5

Monte Carlo Source Identification

5.1 Proofs

Here, we provide proofs of the results in the paper. For integers $m \leq n$, we write $[\![m, n]\!]$ for the integer interval starting at m and ending at n, that is, $\{m, m+1, \ldots, n-1, n\}$.

5.1.1 Proof of Lemma 1

We have

$$\begin{aligned}
\mathbb{E}[\widehat{Y}_{\mathrm{unb}}] &= \mathbb{E}[Y(\widetilde{\sigma})] \\
&= \sum_{\sigma \in \mathrm{Sym}(O)} \mathbb{P}_s^{G_O}(\widetilde{\sigma}^k = \sigma^k) Y(\sigma) \\
&= \sum_{\sigma \in \mathrm{Sym}(O)} \prod_{i=1}^{k-1} \Big[\frac{\partial \mathrm{vol}(\sigma_{[i]}, \sigma_{[i+1]})}{\partial \mathrm{vol}(\sigma_{[i]}, O)} \cdot \frac{\partial \mathrm{vol}(\sigma_{[i]}, O)}{\partial \mathrm{vol}(\sigma_{[i]}, [n])} \Big] \\
&= \sum_{\sigma \in \mathrm{Sym}(O)} \prod_{i=1}^{k-1} \frac{\partial \mathrm{vol}(\sigma_{[i]}, \sigma_{[i+1]})}{\partial \mathrm{vol}(\sigma_{[i]}, [n])} \\
&= \sum_{\sigma \in \mathrm{Sym}(O)} \mathbb{P}_s(\widetilde{\sigma}^k = \sigma^k) = \rho(s \to O),
\end{aligned}$$

which is the desired result.

5.1.2 Proof of Theorem 1

Fix s and O and let us write $\rho(s) := \rho(s \to O)$. We first show that under the ER model for G, the transition probabilities are likely to be exponentially small,

$$\mathbb{P}\Big(\rho(s) \leq (\beta n)^{-k/4}\Big) \geq \frac{1}{2}, \tag{5.1}$$

where the probability is taken w.r.t. $G \sim \mathrm{ER}(n, P)$.

In order to show (5.1), we use the restricted dynamic as a device. Let $\tilde{\sigma} \mid G \sim$ CH-SIP(G_O, s), that is, conditioned on the network being G, $\tilde{\sigma}$ follows the path dynamic restricted to G_O. By Lemma 1, we have $\mathbb{E}[Y(\tilde{\sigma}) \mid G] = \rho(s)$, where $Y(\tilde{\sigma})$ is defined in (2.16). For any $\sigma \in \mathrm{Sym}(O)$, let $q(\sigma) := \mathbb{P}(\tilde{\sigma} = \sigma \mid G)$ as given by (2.17). Note that $\rho(s) = \sum_{\sigma \in \mathrm{Sym}(O)} q(\sigma) Y(\sigma)$. It is then enough to show that $Y(\sigma) \le (\beta n)^{-k/4}$ for all $\sigma \in \mathrm{Sym}(O)$, with probability at least $1/2$.

Fix $U \subset O$ with $|U| = i \in [\![k/4, 3k/4]\!]$. By the Hoeffding inequality,

$$\begin{aligned}\mathbb{P}\Big(\frac{\partial \mathrm{vol}(U,O)}{|U| \cdot |O \setminus U|} \ge p_{\max} + \varepsilon\Big) &\le \exp\Big(-2i(k-i)\varepsilon^2\Big) \\ &\le \exp\Big(-3k^2\varepsilon^2/8\Big)\end{aligned}$$

we have $i(k-i) \ge 3k^2/16$ so the constant is $3/8$ not $1/8$ and

$$\begin{aligned}\mathbb{P}\Big(\frac{\mathrm{vol}(U,O^c)}{|U| \cdot |O^c|} \le p_{\min} - \varepsilon\Big) &\le \exp\Big(-2i(n-k)\varepsilon^2\Big) \\ &\le \exp\Big(-k(n-k)\varepsilon^2/2\Big) \\ &\le \exp(-k^2\varepsilon^2/4)\end{aligned}$$

where the last line uses $k \le \sqrt{n} \le n/2$. It follows that with probability at least $1 - 2\exp(-k^2\varepsilon^2/4)$, we have

$$\frac{\mathrm{vol}(U,O^c)}{\partial \mathrm{vol}(U,O)} \ge \frac{p_{\min} - \varepsilon}{p_{\max} + \varepsilon} \cdot \frac{n-k}{k-i}.$$

We also have

$$\frac{n-k}{k-i} \ge \frac{4}{3}\Big(\frac{n}{k} - 1\Big) \ge \sqrt{n}$$

using $i \ge k/4$, $k \le \sqrt{n}$ and $n \ge 16$. Take $\varepsilon = p_{\min}/2$, and recall the definition of $\beta = p_{\min}^2/(p_{\min} + 2p_{\max})^2$. Then, taking a union bound, we have with probability at least $1 - 2 \cdot 2^k \exp(-k^2 p_{\min}^2/16)$,

$$\frac{\mathrm{vol}(U,O^c)}{\partial \mathrm{vol}(U,O)} \ge \sqrt{\beta n}$$

for all $U \subset O$ with $|U| \in [\![k/4, 3k/4]\!]$. Now, since the vol operator is invariant to the ordering of the element of U, we can apply this inequality with U being the unordered set of elements of $\sigma_{[i]}$ for i in the stated interval. That is, with probability at least $1 - 2^{k+1}\exp(-k^2 p_{\min}^2/16)$, we have

$$\frac{\mathrm{vol}(\sigma_{[i]}, O^c)}{\partial\mathrm{vol}(\sigma_{[i]}, O)} \geq \sqrt{\beta n}$$

for all $\sigma \in \mathrm{Sym}(O)$ and $i \in [\![k/4, 3k/4]\!]$. On this event, we have, for all $\sigma \in \mathrm{Sym}(O)$,

$$\begin{aligned} Y(\sigma) &\leq \prod_{i=k/4}^{3k/4} \frac{\partial\mathrm{vol}(\sigma_{[i]}, O)}{\partial\mathrm{vol}(\sigma_{[i]}, [n])} \\ &= \prod_{i=k/4}^{3k/4} \Big(1 + \frac{\mathrm{vol}(\sigma_{[i]}, O^c)}{\partial\mathrm{vol}(\sigma_{[i]}, O)}\Big)^{-1} \\ &\leq \prod_{i=k/4}^{3k/4} \Big(1 + \sqrt{\beta n}\Big)^{-1} \leq (\beta n)^{-k/4}. \end{aligned}$$

This establishes (5.1) with the desired probability when $(k+2)\log 2 \leq k^2 p_{\min}^2/16$ which holds by assumption (since $k \geq 2$).

Next, we recall that, given G, $\widehat{\rho}_{\mathrm{MC}}^{(N)}$ is the average of N independent Bernoulli variables with probability $\rho(s)$. We thus have

$$\begin{aligned} \mathbb{P}\Big(\widehat{\rho}_{\mathrm{MC}}^{(N)} = 0 \mid \rho(s) \leq (\beta n)^{-k/4}\Big) &\geq \Big(1 - (\beta n)^{-k/4}\Big)^N \\ &\geq 1 - N(\beta n)^{-k/4} \geq 1/2 \end{aligned}$$

using $N \leq (\beta n)^{k/4}/2$. Combining with (5.1) finishes the proof.

5.1.3 Proof of Theorem 2

To simplify the notation, let $\ell(s) := \ell(s \to O)$ and $Z(s) := Z(s \to O)$. Recall that $\ell(s) = -\log \rho(s \to O)$ and note that $\ell(s) = -\log \mathbb{E}[Y(\widetilde{\sigma})]$. We also have

$$\partial\mathrm{vol}(U, W) \leq \min\Big\{|U| d_{\max}, (|W| - |U|)|U|, |W| d_{\max}\Big\}$$

where $d_{\max}$ is the maximum degree of the underlying graph.

Let us write $g(U, W) = \log \partial\mathrm{vol}(U, W)$. Then,

$$g(\widetilde{\sigma}_{[i]}, [n]) \le \log(id_{\max}), \quad g(\widetilde{\sigma}_{[i]}, O) \ge \log(c_2),$$

since $\partial\mathrm{vol}(\widetilde{\sigma}_{[i]}, O) \ge c_2$ by (2.21). We obtain

$$\begin{aligned} -\log Y(\widetilde{\sigma}) &= \sum_{i=1}^{k-1} \left[g(\widetilde{\sigma}_{[i]}, [n]) - g(\widetilde{\sigma}_{[i]}, O) \right] \\ &\le \sum_{i=1}^{k-1} [\log(id_{\max}) - \log(c_2)] \le k \log\left(k \frac{d_{\max}}{c_2} \right). \end{aligned}$$

Next, we obtain a lower bound on $-\log Y(\widetilde{\sigma})$. We have

$$\begin{aligned} \partial\mathrm{vol}(\widetilde{\sigma}_{[i]}, [n]) &= \mathrm{vol}(\widetilde{\sigma}_{[i]}, [n] \setminus \widetilde{\sigma}_{[i]}) \\ &= \mathrm{vol}(\widetilde{\sigma}_{[i]}, O \setminus \widetilde{\sigma}_{[i]}) + \mathrm{vol}(\widetilde{\sigma}_{[i]}, [n] \setminus O). \end{aligned}$$

Then, for any $i \ge k/2 + 1$,

$$\begin{aligned} \frac{\partial\mathrm{vol}(\widetilde{\sigma}_{[i]}, [n])}{\partial\mathrm{vol}(\widetilde{\sigma}_{[i]}, O)} &= 1 + \frac{\mathrm{vol}(\widetilde{\sigma}_{[i]}, [n] \setminus O)}{\partial\mathrm{vol}(\widetilde{\sigma}_{[i]}, O)} \\ &\ge 1 + c_1 \frac{id_{\mathrm{ave}}}{i(d_{\max} \wedge k)} \ge 1 + c_1 \frac{d_{\mathrm{ave}}}{d_{\max}} = e^{\alpha}, \end{aligned}$$

using $\mathrm{vol}(\widetilde{\sigma}_{[i]}, [n] \setminus O) \ge c_1 i\, d_{\mathrm{ave}}$ by (2.20). Hence,

$$Y(\widetilde{\sigma}) \le \prod_{i=1+k/2}^{k} e^{-\alpha} = e^{-(k/2)\alpha}$$

where the first inequality is by dropping the first $k/2$ terms. We obtain $-\log Y(\widetilde{\sigma}) \ge (k/2)\alpha$ and $\ell(s) = -\log \mathbb{E} Y(\widetilde{\sigma}) \ge (k/2)\alpha$. Combining the upper bound on $Y(\widetilde{\sigma})$ with the lower bound on $\ell(s)$, we have

$$X(\widetilde{\sigma}) := \frac{-\log Y(\widetilde{\sigma})}{\ell(s)} \le \frac{2}{\alpha} \log(k d_{\max}/c_2)$$

Since $Y(\widetilde{\sigma}) \le 1$, we also have $X_t(\widetilde{\sigma}) \ge 0$. Recall that a random variable that takes values in $[a, b]$, almost surely, is sub-Gaussian with squared sub-Gaussian parameter $(b - a)^2/4$. Each $X(\widetilde{\sigma}^t)$ is then sub-Gaussian with squared parameter bounded above by $\sigma_1^2 := \log^2(k d_{\max}/c_2)/\alpha^2$. It follows that

$$Z(s \to O) = \left(\frac{1}{N} \sum_{i=1}^{N} X(\widetilde{\sigma}^t) - 1 \right) \sim \mathrm{subG}\left(\frac{\sigma_1^2}{N} \right) \tag{5.2}$$

by the independence of $X(\widetilde{\sigma}^t), t = 1, \ldots, N$ and since the sub-Gaussian parameter is invariant to a constant shift of the variable.

It remains to bound the bias. By Jensen's inequality and a simple lemma

$$0 < \log \mathbb{E}[Y(\widetilde{\sigma})] - \mathbb{E}\Big[\log Y(\widetilde{\sigma})\Big] < \frac{1}{2}\operatorname{var}\Big(Y(\widetilde{\sigma})\Big). \tag{5.3}$$

Since $0 \leq Y(\widetilde{\sigma}) \leq 1$, we have $\mathbb{E}[Y(\widetilde{\sigma})]^2 \leq \mathbb{E}[Y(\widetilde{\sigma})]$, hence

$$\operatorname{var}(Y(\widetilde{\sigma})) \leq \mathbb{E}[Y(\widetilde{\sigma})] = e^{-\ell(s)}$$

Using the variance in equality in (5.3), dividing by $\ell(s)$ and noting $\mathbb{E}[Z(s)]+1 = \mathbb{E}[-\log Y(\widetilde{\sigma})]/\ell(s)$, we obtain

$$0 < \mathbb{E}Z(s \to O) < \frac{1}{h(\ell(s))}.$$

Since $\ell(s)$ is bounded from below by $(k/2)\alpha$, the proof is complete.

5.1.4 Proof of Corollary 1

We begin by writing the weighted Bayes rank loss for $\widehat{\ell}_{\mathrm{IS}}$. Let us abbreviate $\widehat{\ell}_{\mathrm{IS}}(s \to O)$, $\ell(s \to O)$ and $Z(s \to O)$ as $\widehat{\ell}_{\mathrm{IS}}(s)$, $\ell(s)$ and $Z(s)$ respectively. Letting $s^* = \hat{s}_{\mathrm{ML}}(O)$, we have

$$\mathcal{R}_B(\hat{\pi}_{\mathrm{IS}}) = \frac{1}{k}\sum_{s \in O \setminus \{s^*\}} \frac{\ell(s) - \ell(s^*)}{\ell(s)} 1\{\widehat{\ell}_{\mathrm{IS}}(s^*) > \widehat{\ell}_{\mathrm{IS}}(s)\}.$$

For any s for which $\widehat{\ell}_{\mathrm{IS}}(s^*) > \widehat{\ell}_{\mathrm{IS}}(s)$, we have

$$\ell(s) - \ell(s^*) \leq \ell(s) - \widehat{\ell}_{\mathrm{IS}}(s) + \widehat{\ell}_{\mathrm{IS}}(s^*) - \ell(s^*).$$

Recalling that $Z(s) = \widehat{\ell}_{\mathrm{IS}}(s)/\ell(s) - 1$, we get, for any such s,

$$\frac{\ell(s) - \ell(s^*)}{\ell(s)} \leq \frac{\ell(s^*)}{\ell(s)} Z(s^*) - Z(s).$$

We therefore have

$$\begin{aligned} \mathcal{R}_B(\hat{\pi}_{\mathrm{IS}}) \leq \mathcal{Z} &:= \frac{1}{k}\Big(\beta Z(s^*) - \sum_{s \in O \setminus s^*} Z(s)\Big), \quad \text{where} \\ \beta &:= \sum_{s \in O \setminus s^*} \frac{\ell(s^*)}{\ell(s)} \end{aligned} \tag{5.4}$$

Note that $0 < \beta \le k-1$ since $\ell(s^*) \le \ell(s)$.

The estimates $\widehat{\ell}_{\mathrm{IS}}(s)$ are based on independent Monte Carlo samples for different s. It follows that $\{Z(s)\}_{s\in O}$ are independent sub-Gaussian variables, each with parameter

$$\sigma^2 := \sigma_1^2/N = \log^2(kd_{\max}/c_2)/(\alpha^2 N)$$

as established in (5.2). Then, $\mathcal{Z}$ is also sub-Gaussian and

$$\mathcal{Z} \sim \mathrm{subG}\left(\frac{1}{k^2}\Big(\beta^2\sigma^2 + (k-1)\sigma^2\Big)\right) \sim \mathrm{subG}(\sigma^2).$$

Furthermore,

$$\begin{aligned}\mathbb{E}[\mathcal{Z}] &= \frac{\beta}{k}\mathbb{E}[Z(s^*)] - \frac{1}{k}\sum_{s\in O\setminus s^*}\mathbb{E}[Z(s)] \\ &\le \mathbb{E}[Z(s^*)] \;\le\; \frac{1}{h(k\alpha/2)}\end{aligned}$$

using $\beta \le k-1$ and $\mathbb{E}Z(s) > 0$, for all s, to obtain the first inequality.

Putting the pieces together and using sub-Gaussian concentration, we get

$$\begin{aligned}\mathbb{P}\Big(\mathcal{R}_B(\hat{\pi}_{\mathrm{IS}}) > \sigma t + \frac{1}{h(\alpha k/2)}\Big) &\le \mathbb{P}\Big(\mathcal{Z} > \sigma t + \mathbb{E}[\mathcal{Z}]\Big) \\ &\le \exp(-t^2/2)\end{aligned}$$

Taking $t = \sqrt{2\log(1/\delta)}$ finishes the proof.

5.1.5 Auxiliary lemmas

Lemma 3. *Assume that Y is a random variable such that $-A \le \log Y \le 0$ almost surely. Then,*

$$\frac{1}{2}\operatorname{var}(Y) \le \log(\mathbb{E}Y) - \mathbb{E}[\log Y] \le \frac{1}{2}e^{2A}\operatorname{var}(Y).$$

Proof. Let $f(x) = \log x$ and $\mu = \mathbb{E}Y$. By Taylor expansion,

$$f(Y) = f(\mu) + f'(\mu)(Y-\mu) + \frac{1}{2}f''(\tilde{\mu})(Y-\mu)^2.$$

where $\widetilde{\mu}$ is between Y and μ. Taking expectation and rearranging

$$f(\mu) - \mathbb{E}[f(Y)] = \frac{1}{2}\mathbb{E}[-f''(\widetilde{\mu})(Y-\mu)^2].$$

We have $\widetilde{\mu} \in [e^{-A}, 1]$, hence $-f''(\widetilde{\mu}) = 1/\widetilde{\mu}^2 \in [1, e^{2A}]$. The result follows. □

Lemma 4. *Let Z be a bounded random variable whose third central moment is negative. Let $v = \mathbb{E}Z$. Then,*

$$-v + \log \mathbb{E}[e^Z] \le \frac{1}{2}\operatorname{var}(Z).$$

Proof. By the Taylor expansion,

$$e^Z = e^v + e^v(Z-v) + \frac{e^v}{2}(Z-v)^2 + \frac{e^{\tilde{v}}}{3!}(Z-v)^3$$

where $\tilde{v}$ is between Z and v. Assuming that $|Z| \le b$, then $|v - \tilde{v}| \le b$ and

$$\begin{aligned} e^{-v}\mathbb{E}[e^Z] &= 1 + \frac{1}{2}\operatorname{var}(Z) + \frac{1}{3!}\mathbb{E}[e^{\tilde{v}-v}(Z-v)^3] \\ &\le 1 + \frac{1}{2}\operatorname{var}(Z) + \frac{1}{3!}e^b\mathbb{E}[(Z-v)^3] \\ &\le 1 + \frac{1}{2}\operatorname{var}(Z) \end{aligned}$$

where the last line is by negative skewness. Taking log of both sides and using $\log(1+x) \le x$, gives the result. □

We can apply the lemma to $Z = \log Y(\widetilde{\sigma})$. Note that $-A \le Z \le -B$ where $A = k\log(kd_{\max}/c_2)$ and $B = (k/2)\alpha$. Then since $\log \mathbb{E}[e^Z] = -\ell(s)$, we have

$$\frac{-v}{\ell(s)} - 1 \le \frac{1}{2}\frac{\operatorname{var}(Z)}{\ell(s)} \lesssim \frac{\sigma_1^2}{(k/2)\alpha} \lesssim \frac{\log^2(kd_{\max}/c_2)}{k\alpha^3}$$

which is $o(1)$ as $k \to 0$.

[AAM92] Roy M Anderson, B Anderson, and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

[AAR10] Dana Angluin, James Aspnes, and Lev Reyzin. "Inferring social networks from outbreaks." In *International conference on algorithmic learning theory*, pp. 104–118. Springer, 2010.

[ABD14] Fabrizio Altarelli, Alfredo Braunstein, Luca Dall'Asta, Alejandro Lage-Castellanos, and Riccardo Zecchina. "Bayesian inference of epidemics on networks via belief propagation." *Physical review letters*, **112**(11):118701, 2014.

[AG17] Hunt Allcott and Matthew Gentzkow. "Social media and fake news in the 2016 election." *Journal of economic perspectives*, **31**(2):211–36, 2017.

[ALS15] Nino Antulov-Fantulin, Alen Lančić, Tomislav Šmuc, Hrvoje Štefančić, and Mile Šikić. "Identification of patient zero in static and temporal networks: Robustness and limitations." *Physical review letters*, **114**(24):248701, 2015.

[AOT15] Daron Acemoglu, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. "Systemic risk and stability in financial networks." *American Economic Review*, **105**(2):564–608, 2015.

[BB04] James M Bower and Hamid Bolouri. *Computational modeling of genetic and biochemical networks*. MIT press, 2004.

[BFV17] Shaileshh Bojja Venkatakrishnan, Giulia Fanti, and Pramod Viswanath. "Dandelion: Redesigning the bitcoin network for anonymity." *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, **1**(1):22, 2017.

[BKM17] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians." *Journal of the American statistical Association*, **112**(518):859–877, 2017.

[CBB06] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. "The role of the airline transportation network in the prediction and predictability of global epidemics." *Proceedings of the National Academy of Sciences*, **103**(7):2015–2020, 2006.

[CH04] Andrew Cliff and Peter Haggett. "Time, travel and infection." *British medical bulletin*, **69**(1):87–99, 2004.

[Coh00] Mitchell L Cohen. "Changing patterns of infectious disease." *Nature*, **406**(6797):762, 2000.

[CZC15a] Biao Chang, Feida Zhu, Enhong Chen, and Qi Liu. "Information source detection via maximum a posteriori estimation." In *Data Mining (ICDM), 2015 IEEE International Conference on*, pp. 21–30. IEEE, 2015.

[CZC15b] Biao Chang, Feida Zhu, Enhong Chen, and Qi Liu. "Information source detection via maximum a posteriori estimation." In *2015 IEEE International Conference on Data Mining*, pp. 21–30. IEEE, 2015.

[DCB17] Gytis Dudas, Luiz Max Carvalho, Trevor Bedford, Andrew J Tatem, Guy Baele, Nuno R Faria, Daniel J Park, Jason T Ladner, Armando Arias, Danny Asogun, et al. "Virus genomes reveal factors that spread and sustained the Ebola epidemic." *Nature*, **544**(7650):309, 2017.

[DPH12] Wen Dong, Alex Pentland, and Katherine A Heller. "Graph-coupled hmms for modeling the spread of infection." *arXiv preprint arXiv:1210.4864*, 2012.

[EGJ14] Matthew Elliott, Benjamin Golub, and Matthew O Jackson. "Financial networks and contagion." *American Economic Review*, **104**(10):3115–53, 2014.

[EK10] David Easley, Jon Kleinberg, et al. *Networks, crowds, and markets*, volume 8. Cambridge university press Cambridge, 2010.

[FAE14] Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. "Rumor cascades." In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[FC12] Vincenzo Fioriti and Marta Chinnici. "Predicting the sources of an outbreak with a spectral technique." *arXiv preprint arXiv:1211.2333*, 2012.

[FCP14] Vincenzo Fioriti, Marta Chinnici, and Jesus Palomo. "Predicting the sources of an outbreak with a spectral technique." *Applied Mathematical Sciences*, **8**(133-136):6775–6782, 2014.

[FHL15] Ling Feng, Yanqing Hu, Baowen Li, H Eugene Stanley, Shlomo Havlin, and Lidia A Braunstein. "Competing for attention in social media under information overload conditions." *PloS one*, **10**(7):e0126090, 2015.

[FLJ07] Chris Fleizach, Michael Liljenstam, Per Johansson, Geoffrey M Voelker, and Andras Mehes. "Can you infect me now?: malware propagation in mobile phone networks." In *Proceedings of the 2007 ACM workshop on Recurring malcode*, pp. 61–68. ACM, 2007.

[Gew89] John Geweke. "Bayesian inference in econometric models using Monte Carlo integration." *Econometrica: Journal of the Econometric Society*, pp. 1317–1339, 1989.

[GLK12] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. "Inferring networks of diffusion and influence." *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **5**(4):21, 2012.

[GSD16] Manuel Gomez-Rodriguez, Le Song, Hadi Daneshmand, and Bernhard Schölkopf. "Estimating diffusion networks: Recovery conditions, sample complexity & soft-thresholding algorithm." *The Journal of Machine Learning Research*, **17**(1):3092–3120, 2016.

[HF19] Abigail L Horn and Hanno Friedrich. "Locating the source of large-scale outbreaks of foodborne disease." *Journal of the Royal Society Interface*, **16**(151):20180624, 2019.

[JCG17] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, and Jiebo Luo. "Detection and analysis of 2016 us presidential election related rumors on twitter." In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pp. 14–24. Springer, 2017.

[JS08] Björn H Junker and Falk Schreiber. *Analysis of biological networks*, volume 2. Wiley Online Library, 2008.

[JWY16] Jiaojiao Jiang, Sheng Wen, Shui Yu, Yang Xiang, and Wanlei Zhou. "Identifying propagation sources in networks: State-of-the-art and comparative studies." *IEEE Communications Surveys & Tutorials*, **19**(1):465–481, 2016.

[JWY17] Jiaojiao Jiang, Sheng Wen, Shui Yu, Yang Xiang, and Wanlei Zhou. "Identifying propagation sources in networks: State-of-the-art and comparative studies." *IEEE Communications Surveys & Tutorials*, **19**(1):465–481, 2017.

[KA19] S Jalil Kazemitabar and Arash A Amini. "Approximate Identification of the Optimal Epidemic Source in Complex Networks." *arXiv preprint arXiv:1906.03052*, 2019.

[KGH10] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. "Identification of influential spreaders in complex networks." *Nature physics*, **6**(11):888, 2010.

[KKT03] David Kempe, Jon Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network." In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146. ACM, 2003.

[KL17a] Justin Khim and Po-Ling Loh. "Confidence sets for the source of a diffusion in regular trees." *IEEE Transactions on Network Science and Engineering*, **4**(1):27–40, 2017.

[KL17b] Justin Khim and Po-Ling Loh. "Permutation tests for infection graphs." *arXiv preprint arXiv:1705.07997*, 2017.

[KL18] Justin Khim and Po-Ling Loh. "A theory of maximum likelihood for weighted infection graphs." *arXiv preprint arXiv:1806.05273*, 2018.

[KM27] William Ogilvy Kermack and Anderson G McKendrick. "A contribution to the mathematical theory of epidemics." *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, **115**(772):700–721, 1927.

[KMS17] István Z Kiss, Joel C Miller, Péter L Simon, et al. "Mathematics of epidemics on networks." *Cham: Springer*, **598**, 2017.

[KN11] Brian Karrer and Mark EJ Newman. "Stochastic blockmodels and community structure in networks." *Physical review E*, **83**(1):016107, 2011.

[Kon08] Suleyman Kondakci. "Epidemic state analysis of computers under malware attacks." *Simulation Modelling Practice and Theory*, **16**(5):571–584, 2008.

[KS] J Komlós and M Simonovits. "Szemerédi's regularity lemma and its applications in graph theory. Combinatorics, Paul Erdos is eighty, Vol. 2 (Keszthely, 1993), 295–352." *Bolyai Soc. Math. Stud*, **2**.

[LDP15] A Lima, M De Domenico, V Pejovic, and M Musolesi. "Disease containment strategies based on mobility and information dissemination." *Scientific reports*, **5**:10650, 2015.

[LHK10] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. "Signed networks in social media." In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1361–1370. ACM, 2010.

[LKF05] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. "Graphs over time: densification laws, shrinking diameters and possible explanations." In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 177–187. ACM, 2005.

[LMO14] Andrey Y Lokhov, Marc Mézard, Hiroki Ohta, and Lenka Zdeborová. "Inferring the origin of an epidemic with a dynamic message-passing algorithm." *Physical Review E*, **90**(1):012801, 2014.

[LRB14] Philippe Lemey, Andrew Rambaut, Trevor Bedford, Nuno Faria, Filip Bielejec, Guy Baele, Colin A Russell, Derek J Smith, Oliver G Pybus, Dirk Brockmann, et al. "Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2." *PLoS pathogens*, **10**(2):e1003932, 2014.

[LT12] Wuqiong Luo and Wee Peng Tay. "Identifying multiple infection sources in a network." In *Signals, Systems and Computers (ASILOMAR), 2012 Conference Record of the Forty Sixth Asilomar Conference on*, pp. 1483–1489. IEEE, 2012.

[LTG10] Theodoros Lappas, Evimaria Terzi, Dimitrios Gunopulos, and Heikki Mannila. "Finding effectors in social networks." In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1059–1068. ACM, 2010.

[LTL14] Wuqiong Luo, Wee Peng Tay, and Mei Leng. "How to identify an infection source with limited observations." *IEEE Journal of Selected Topics in Signal Processing*, **8**(4):586–597, 2014.

[LWG17] Mei Li, Xiang Wang, Kai Gao, and Shanshan Zhang. "A survey on information diffusion in online social networks: Models and methods." *Information*, **8**(4):118, 2017.

[MCM15] Chris Milling, Constantine Caramanis, Shie Mannor, and Sanjay Shakkottai. "Distinguishing infections on different graph topologies." *IEEE Transactions on Information Theory*, **61**(6):3100–3120, 2015.

[MKS14a] Juliane Manitz, Thomas Kneib, Martin Schlather, Dirk Helbing, and Dirk Brockmann. "Origin detection during food-borne disease outbreaks-a case study of the 2011 ehec/hus outbreak in germany." *PLoS currents*, **6**, 2014.

[MKS14b] Juliane Manitz, Thomas Kneib, Martin Schlather, Dirk Helbing, and Dirk Brockmann. "Origin detection during food-borne disease outbreaks-a case study of the 2011 ehec/hus outbreak in germany." *PLoS currents*, **6**, 2014.

[NGM16a] Hung T Nguyen, Preetam Ghosh, Michael L Mayo, and Thang N Dinh. "Multiple infection sources identification with provable guarantees." In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 1663–1672. ACM, 2016.

[NGM16b] Hung T Nguyen, Preetam Ghosh, Michael L Mayo, and Thang N Dinh. "Multiple infection sources identification with provable guarantees." In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 1663–1672. ACM, 2016.

[NOS12] Sahand Negahban, Sewoong Oh, and Devavrat Shah. "Iterative ranking from pair-wise comparisons." In *Advances in neural information processing systems*, pp. 2474–2482, 2012.

[Ore] U. of Oregon Route Views Project. "Online data and reports." `http://www.routeviews.org`.

[Org08] World Health Organization. *Foodborne disease outbreaks: guidelines for investigation and control*. World Health Organization, 2008.

[OUS08] Avi Ostfeld, James G Uber, Elad Salomons, Jonathan W Berry, William E Hart, Cindy A Phillips, Jean-Paul Watson, Gianluca Dorini, Philip Jonkergouw, Zoran Kapelan, et al. "The battle of the water sensor networks (BWSN): A design challenge for engineers and algorithms." *Journal of Water Resources Planning and Management*, **134**(6):556–568, 2008.

[PJZ19] Xin Pei, Zhen Jin, Wenyi Zhang, and Yong Wang. "Detection of Infection Sources for Avian Influenza A (H7N9) in Live Poultry Transport Network During the Fifth Wave in China." *IEEE Access*, **7**:155759–155778, 2019.

[PLS18] Robert Paluch, Xiaoyan Lu, Krzysztof Suchecki, Bolesław K Szymański, and Janusz A Hołyst. "Fast and accurate detection of spread source in large complex networks." *Scientific reports*, **8**(1):2508, 2018.

[PMA14] Sen Pei, Lev Muchnik, José S Andrade Jr, Zhiming Zheng, and Hernán A Makse. "Searching for superspreaders of information in real-world social media." *Scientific reports*, **4**:5547, 2014.

[PV18] Bastian Prasse and Piet Van Mieghem. "Maximum-likelihood network reconstruction for SIS processes is NP-hard." *arXiv preprint arXiv:1807.08630*, 2018.

[PVF12] B Aditya Prakash, Jilles Vreeken, and Christos Faloutsos. "Spotting culprits in epidemics: How many and which ones?" In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pp. 11–20. IEEE, 2012.

[RBP13] Tarik Roukny, Hugues Bersini, Hugues Pirotte, Guido Caldarelli, and Stefano Battiston. "Default cascades in complex networks: Topology and systemic risk." *Scientific reports*, **3**:2759, 2013.

[Ros00] Lewis A Rossman et al. "EPANET 2: users manual." 2000.

[Ros14] Sheldon M Ross. *Introduction to probability models*. Academic press, 2014.

[SAS98] Laurence Slutsker, Sean F Altekruse, and David L Swerdlow. "Foodborne diseases: emerging pathogens and trends." *Infectious Disease Clinics*, **12**(1):199–216, 1998.

[SCV17] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. "The spread of fake news by social bots." *arXiv preprint arXiv:1707.07592*, pp. 96–104, 2017.

[SCV18] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. "The spread of low-credibility content by social bots." *Nature communications*, **9**(1):4787, 2018.

[SCW16] Zhesi Shen, Shinan Cao, Wen-Xu Wang, Zengru Di, and H Eugene Stanley. "Locating the source of diffusion in complex networks by time-reversal backward spreading." *Physical Review E*, **93**(3):032301, 2016.

[SJD17] Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. "Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction." *new media & society*, **19**(8):1214–1235, 2017.

[Sni02] Tom AB Snijders. "Markov chain Monte Carlo estimation of exponential random graph models." *Journal of Social Structure*, **3**(2):1–40, 2002.

[SR05] Vishal Sood and Sidney Redner. "Voter model on heterogeneous graphs." *Physical review letters*, **94**(17):178701, 2005.

[SZ11a] Devavrat Shah and Tauhid Zaman. "Rumors in a network: Who's the culprit?" *IEEE Transactions on information theory*, **57**(8):5163–5181, 2011.

[SZ11b] Devavrat Shah and Tauhid Zaman. "Rumors in a network: Who's the culprit?" *IEEE Transactions on information theory*, **57**(8):5163–5181, 2011.

[TKM11] Amanda L Traud, Eric D Kelsic, Peter J Mucha, and Mason A Porter. "Comparing community structure to characteristics in online collegiate social networks." *SIAM review*, **53**(3):526–543, 2011.

[TMP12] Amanda L Traud, Peter J Mucha, and Mason A Porter. "Social structure of Facebook networks." *Physica A: Statistical Mechanics and its Applications*, **391**(16):4165–4180, 2012.

[WJ08] Martin J Wainwright, Michael I Jordan, et al. "Graphical models, exponential families, and variational inference." *Foundations and Trends® in Machine Learning*, **1**(1–2):1–305, 2008.

[WLJ15] Qiyao Wang, Zhen Lin, Yuehui Jin, Shiduan Cheng, and Tan Yang. "ESIS: emotion-based spreader–ignorant–stifler model for information diffusion." *Knowledge-Based Systems*, **81**:46–55, 2015.

[WS98] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world'networks." *Nature*, **393**(6684):440, 1998.

[ZY16a] Kai Zhu and Lei Ying. "Information source detection in the SIR model: A sample-path-based approach." *IEEE/ACM Transactions on Networking (TON)*, **24**(1):408–421, 2016.

[ZY16b] Kai Zhu and Lei Ying. "Information source detection in the SIR model: A sample-path-based approach." *IEEE/ACM Transactions on Networking (TON)*, **24**(1):408–421, 2016.