Sankhyā : The Indian Journal of Statistics
2022, Volume 84-A, Part 1, pp. 321-344

Scalable Estimation of Epidemic Thresholds via Node Sampling

Anirban Dasgupta
Indian Institute of Technology, Gandhinagar, Gandhinagar, India
Srijan Sengupta
North Carolina State University, Raleigh, USA

Abstract

Infectious or contagious diseases can be transmitted from one person to another through social contact networks. In today's interconnected global society, such contagion processes can cause global public health hazards, as exemplified by the ongoing Covid-19 pandemic. It is therefore of great practical relevance to investigate the network transmission of contagious diseases from the perspective of statistical inference. An important and widely studied boundary condition for contagion processes over networks is the so-called *epidemic threshold*. The epidemic threshold plays a key role in determining whether a pathogen introduced into a social contact network will cause an epidemic or die out. In this paper, we investigate epidemic thresholds from the perspective of statistical network inference. We identify two major challenges that are caused by high computational and sampling complexity of the epidemic threshold. We develop two statistically accurate and computationally efficient approximation techniques to address these issues under the Chung-Lu modeling framework. The second approximation, which is based on random walk sampling, further enjoys the advantage of requiring data on a vanishingly small fraction of nodes. We establish theoretical guarantees for both methods and demonstrate their empirical superiority.

AMS (2000) *subject classification.* 62F10 (primary), 68W20, 68W25.
Keywords and phrases. Epidemic threshold, Networks, Sampling, Random walk, Configuration model, Epidemiology.

Anirban Dasgupta's work is partially supported by grants from DBT India, Google and CISCO.
Srijan Sengupta's work is partially supported by an NIH R01 grant 1R01LM013309.

1 Introduction

Infectious diseases are caused by pathogens, such as bacteria, viruses, fungi, and parasites. Many infectious diseases are also contagious, which means the infection can be transmitted from one person to another when there is some interaction (e.g., physical proximity) between them. Today, we live in an interconnected world where such contagious diseases could spread through social contact networks to become global public health hazards. A recent example of this phenomenon is the Covid-19 outbreak caused by the so-called novel coronavirus (SARS-CoV-2) that has spread to many countries (Huang et al., 2020; Zhu et al., 2020; Wang et al., 2020; Sun et al., 2020). This recent global outbreak has caused serious social and economic repercussions, such as massive restrictions on movement and share market decline (Chinazzi et al., 2020). It is therefore of great practical relevance to investigate the transmission of contagious diseases through social contact networks from the perspective of statistical inference.

Consider an infection being transmitted through a population of n individuals. According to the susceptible-infected-recovered (SIR) model of disease spread, the pathogen can be transmitted from an infected person (I) to a susceptible person (S) with an infection rate given by β, and an infected individual becomes recovered (R) with a recovery rate given by μ. This can be modeled as a Markov chain whose state at time t is given by a vector $(X_1^t, \ldots, X_n^t)$, where X_i^t denotes the state of the i^{th} individual at time t, i.e., $X_i^t \in \{S, I, R\}$. For the population of n individuals, the state space of this Markov chain becomes extremely large with 3^n possible configurations, which makes it impractical to study the exact system. This problem was addressed in a series of three seminal papers by Kermack and McKendrick (1927, 1932, 1933). Instead of modeling the disease state of each individual at at a given point of time, they proposed compartmental models, where the goal is to model the number of individuals in a particular disease state (e.g., susceptible, infected, recovered) at a given point of time. Since their classical papers, there has been a tremendous amount of work on compartmental modeling of contagious diseases over the last ninety years (Hethcote, 2000; Van den Driessche and Watmough, 2002; Brauer and Castillo-Chavez, 2012).

Compartmental models make the unrealistic assumption of homogeneity, i.e., each individual is assumed to have the same probability of interacting with any other individual. In reality, individuals interact with each other in a highly heterogeneous manner, depending upon various factors such as age, cultural norms, lifestyle, weather, etc. The contagion process can be significantly impacted by heterogeneity of interactions (Meyers et al., 2005; Rocha et al., 2011; Galvani and May, 2005; Woolhouse et al., 1997), and

therefore compartmental modeling of contagious diseases can lead to substantial errors.

In recent years, contact networks have emerged as a preferred alternative to compartmental models (Keeling, 2005). Here, a node represents an individual, and an edge between two nodes represent social contact between them. An edge connecting an infected node and a susceptible node represents a potential path for pathogen transmission. This framework can realistically represent the heterogeneous nature of social contacts, and therefore provide much more accurate modeling of the contagion process than compartmental models. Notable examples where the use of contact networks have led to improvements in prediction or understanding of infectious diseases include Bengtsson et al. (2015) and Kramer et al. (2016).

Consider the scenario where a pathogen is introduced into a social contact network and it spreads according to an SIR model. It is of particular interest to know whether the pathogen will die out or lead to an epidemic. This is dictated by a set of boundary conditions known as the *epidemic threshold*, which depends on the SIR parameters β and μ as well as the network structure itself. Above the epidemic threshold, the pathogen invades and infects a finite fraction of the population. Below the epidemic threshold, the prevalence (total number of infected individuals) remains infinitesimally small in the limit of large networks (Pastor-Satorras et al., 2015). There is growing evidence that such thresholds exist in real-world host-pathogen systems, and intervention strategies are formulated and executed based on estimates of the epidemic threshold. (Dallas et al., 2018; Shulgin et al., 1998; Wallinga et al., 2005; Pourbohloul et al., 2005; Meyers et al., 2005). Fittingly, the last two decades have seen a significant emphasis on studying epidemic thresholds of contact networks from several disciplines, such as computer science, physics, and epidemiology (Newman 2002; Wang et al. 2003; Colizza and Vespignani 2007; Chakrabarti et al. 2008; Gómez et al. 2010; Wang et al. 2016, 2017). See Leitch et al. (2019) for a complete survey on the topic of epidemic thresholds.

Concurrently but separately, network data has rapidly emerged as a significant area in statistics. Over the last two decades, a substantial amount of methodological advancement has been accomplished in several topics in this area, such as community detection (Bickel and Chen, 2009; Zhao et al., 2012; Rohe et al., 2011; Sengupta and Chen, 2015), model fitting and model selection (Hoff et al., 2002; Handcock et al., 2007; Krivitsky et al., 2009; Wang and Bickel, 2017; Yan et al., 2014; Bickel and Sarkar, 2016; Sengupta and Chen, 2018), hypothesis testing (Ghoshdastidar and von Luxburg 2018; Tang et al. 2017a, b; Bhadra et al. 2019), and anomaly detection (Zhao

et al., 2018; Sengupta, 2018; Komolafe et al., 2019), to name a few. The state-of-the-art toolbox of statistical network inference includes a range of random graph models and a suite of estimation and inference techniques.

However, there has not been any work at the intersection of these two areas, in the sense that the problem of estimating epidemic thresholds has not been investigated from the perspective of statistical network inference. Furthermore, the task of computing the epidemic threshold based on existing results can be computationally infeasible for massive networks. In this paper, we address these gaps by developing a novel sampling-based method to estimate the epidemic threshold under the widely used Chung-Lu model (Aiello et al., 2000), also known as the configuration model. We prove that our proposed method has theoretical guarantees for both statistical accuracy and computational efficiency. We also provide empirical results demonstrating our method on both synthetic and real-world networks.

The rest of the paper is organized as follows. In Section 2, we formally set up the problem statement and formulate our proposed methods for approximating the epidemic threshold. In Section 3, we describe the theoretical properties of our estimators. In Section 4, we report numerical results from synthetic as well as real-world networks. We conclude the paper with discussion and next steps in Section 5.

2 Epidemic Thresholds

Table 1 lists the common symbols used throughout the paper. Consider a set of n individuals labelled as $1, \ldots, n$, and an undirected network (with no self-loops) representing interactions between them. This can represented by

Table 1: Common symbols

Symbol	Definition and description
$\lambda(\mathbf{A})$	Spectral radius of the matrix $\mathbf{A}$
d_i	Degree of the node i of the network
δ_i	Expected degree of the node i of the network
$S(t), I(t), R(t)$	Number of susceptible (S), infected (I), and recovered/ removed (R) individuals in the population at time t
β	Infection rate: probability of transmission of a pathogen from an infected individual to a susceptible individual per effective contact (e.g. contact per unit time in continuous-time models, or per time step in discrete-time models)
μ	Recovery rate: probability that an infected individual will recover per unit time (in continuous-time models) or per time step (in discrete-time models)

an n-by-n symmetric adjacency matrix A, where $A(i,j) = 1$ if individuals i and j interact and $A(i,j) = 0$ otherwise. Consider a pathogen spreading through this contact network according to an SIR model. From existing work (Chakrabarti et al. 2008; Gómez et al. 2010; Prakash et al. 2010; Wang et al. 2016, 2017), we know that the boundary condition for the pathogen to become an epidemic is given by

$$\frac{\beta}{\mu} = \frac{1}{\lambda(A)}, \tag{2.1}$$

where $\lambda(A)$ is the spectral radius of the adjacency matrix A.

The left hand side of Eq. 2.1 is the ratio of the infection rate to the recovery rate, which is purely a function of the pathogen and independent of the network. As this ratio grows larger, an epidemic becomes more likely, as new infections outpace recoveries. The right hand side of Eq. 2.1 is the spectral radius of the adjacency matrix, which is purely a function of the network and independent of the pathogen. Larger the spectral radius, the more connected the network, and therefore an epidemic becomes more likely. Thus, the boundary condition in Eq. 2.1 connects the two aspects of the contagion process, the pathogen transmissibility which is quantified by β/μ, and the social contact network which is quantified by the spectral radius. If $\frac{\beta}{\mu} < \frac{1}{\lambda(A)}$, the pathogen dies out, and if $\frac{\beta}{\mu} > \frac{1}{\lambda(A)}$, the pathogen becomes an epidemic.

Given a social contact network, the inverse of the spectral radius of its adjacency matrix represents the epidemic threshold for the network. Any pathogen whose transmissiblity ratio is greater than this threshold is going to cause an epidemic, whereas any pathogen whose transmissiblity ratio is less than this threshold is going to die out. Therefore, a key problem in network epidemiology is to compute the spectral radius of the social contact network.

2.1. Problem Statement and Heuristics Realistic urban social networks that are used in modeling contagion processes have millions of nodes (Eubank et al., 2004; Barrett et al., 2008). To compute the epidemic threshold of such networks, we need to find the largest (in absolute value) eigenvalue of the adjacency matrix A. This is challenging because of two reasons.

1. First, from a computational perspective, eigenvalue algorithms have computational complexity of $\Omega(n^2)$ or higher. For massive social contact networks with millions of nodes, this can become too burdensome.

2. Second, from a statistical perspective, eigenvalue algorithms require the entire adjacency matrix for the full network of n individuals. It can

be challenging or expensive to collect interaction data of n individuals of a massive population (e.g., an urban metropolis). Furthermore, eigenvalue algorithms typically require the full matrix to be stored in the random-access memory of the computer, which can be infeasible for massive social contact networks which are too large to be stored.

The first issue could be resolved if we could compute the epidemic threshold in a computationally efficient manner. The second issue could be resolved if we could compute the epidemic threshold only using data on a small subset of the population. In this paper, we aim to resolve both issues by developing two approximation methods for computing the spectral radius.

To address these problems, let us look at the spectral radius, $\lambda(A)$, from the perspective of random graph models. The statistical model is given by $A \sim P$, which is short-hand for $A(i,j) \sim \text{Bernoulli}(P(i,j))$ for $1 \leq i < j \leq n$. Then $\lambda(A)$ converges to $\lambda(P)$ in probability under some mild conditions (Chung and Radcliffe, 2011; Benaych-Georges et al., 2019; Bordenave et al., 2020). To make a formal statement regarding this convergence, we reproduce below a slightly paraphrased version (for notational consistency) of an existing result in this context.

Lemma 1 (Theorem 1 of Chung and Radcliffe (2011)). *Let*

$$\Delta = \max_{1 \leq i \leq n} \sum_{j=1}^{n} P(i,j)$$

be the maximum expected degree, and suppose that for some $\epsilon > 0$,

$$\Delta > \frac{4}{9} \log(2n/\epsilon)$$

for sufficiently large n. *Then with probability at least* $1 - \epsilon$, *for sufficiently large* n,

$$|\lambda(A) - \lambda(P)| \leq 2\sqrt{\Delta \log(2n/\epsilon)}.$$

To make note of a somewhat subtle point: from an inferential perspective it is tempting to view the above result as a consistency result, where $\lambda(P)$ is the population quantity or parameter of interest and $\lambda(A)$ is its estimator. However, in the context of epidemic thresholds, we are interested in the random variable $\lambda(A)$ itself, as we want to study the contagion spread conditional on a given social contact network. Therefore, in the present context, the above result should not be interpreted as a consistency result.

Rather, we can use the convergence result in a different way. For massive networks, the random variable $\lambda(A)$, which we wish to compute but find it

infeasible to do so, is close to the parameter $\lambda(P)$. Suppose we can find a random variable $T(A)$ which also converges in probability to $\lambda(P)$, and is computationally efficient since $T(A)$ and $\lambda(A)$ both converge in probability to $\lambda(P)$, we can use $T(A)$ as an accurate proxy for $\lambda(A)$. This would address the first of the two issues described at the beginning of this subsection. Furthermore, if $T(A)$ can be computed from a small subset of the data, that would also solve the second issue. This is our central heuristic, which we are going to formalize next.

2.2. The Chung-Lu Model So far, we have not made any structural assumptions on P, we have simply considered the generic inhomogeneous random graph model. Under such a general model, it is very difficult to formulate a statistic $T(A)$ which is cheap to compute and converges to $\lambda(P)$. Therefore, we now introduce a structural assumption on P, in the form of the well-known Chung-Lu model that was introduced by Aiello et al. (2000) and subsequently studied in many papers (Chung and Lu, 2002; Chung et al., 2003; Decreusefond et al., 2012; Pinar et al., 2012; Zhang et al., 2017). For a network with n nodes, let $\delta = (\delta_1, \ldots, \delta_n)'$ be the vector of expected degrees. Then under the Chung-Lu model,

$$P(i,j) = \frac{\delta_i \delta_j}{\sum_{k=1}^{n} \delta_k}. \tag{2.2}$$

This formulation preserves $E[d_i] = \delta_i$, where d_i is the degree of the i^{th} node, and is very flexible with respect to degree heterogeneity.

Under model Eq. 2.2, note that $rank(P) = 1$, and we have

$$\begin{aligned} & P = \frac{1}{\sum_{i=1}^{n} \delta_i} \delta\delta' \\ \Rightarrow \quad & P\delta = \frac{1}{\sum_{i=1}^{n} \delta_i} \delta\delta'\delta = \frac{\sum_{i=1}^{n} \delta_i^2}{\sum_{i=1}^{n} \delta_i} \delta \\ \Rightarrow \quad & \lambda(P) = \frac{\sum_{i=1}^{n} \delta_i^2}{\sum_{i=1}^{n} \delta_i}. \end{aligned}$$

Recall that we are looking for some computationally efficient $T(A)$ which converges in probability to $\lambda(P)$. We now know that under the Chung-Lu model, $\lambda(P)$ is equal to the ratio of the second moment to the first moment of the degree distribution. Therefore, a simple estimator of $\lambda(P)$ is given by the sample analogue of this ratio, i.e.,

$$T_1(A) = \frac{\sum_{i=1}^{n} d_i^2}{\sum_{i=1}^{n} d_i}. \tag{2.3}$$

We now want to demonstrate that approximating $\lambda(A)$ by $T_1(A)$ provides us with very substantial computational savings with little loss of accuracy. The approximation error can be quantified as

$$e_1(A) = \left| \frac{T_1(A)}{\lambda(A)} - 1 \right|, \tag{2.4}$$

and our goal is to show that $e_1(A) \to 0$ in probability, while the computational cost of $T_1(A)$ is much smaller than that of $\lambda(A)$. We will show this both from a theoretical perspective and an empirical perspective. We next describe the empirical results from a simulation study, and we postpone the theoretical discussion to Section 3 for organizational clarity.

We used $n = 5000$, and constructed a Chung-Lu random graph model where $P(i, j) = \theta_i \theta_j$. The model parameters $\theta_1, \ldots, \theta_n$ determine the expected degrees. We used two models for generating θ_i. In the Uniform model, θ_i were uniformly sampled from $(0, 0.25)$. In the PowerLaw model, θ_i were uniformly sampled from the PowerLaw distribution with parameters $x_{min} = 0.01$, $\beta = 3$. Note that the second model leads to heavy-tailed distribution.

Then, we randomly generated 20 networks from the model, and computed $\lambda(A)$ and $T_1(A)$. The results are reported in Table 2. We observe that the runtimes for $T_1(A)$ are orders of magnitude faster than computing the eigenvalue. The average error for $T_1(A)$ is small, and so is the standard deviation (SD) of errors. Thus, even for moderately sized networks, using $T_1(A)$ as a proxy for $\lambda(A)$ can reduce the computational cost to a great extent, without much loss in accuracy. For massive networks where n is in millions, this advantage of $T_1(A)$ over $\lambda(A)$ is even greater; however, the computational burden for $\lambda(A)$ becomes so large that this case is difficult to illustrate using standard computing equipment.

Thus, $T_1(A)$ provides us with a computationally efficient and statistically accurate method for finding the epidemic threshold.

Comparing the results from Uniform and PowerLaw, we observe that errors are higher for the PowerLaw model. A likely explanation for this is that since the distribution is heavy tailed, the moment based estimator is less accurate. This is particularly true for larger n, since the impact of extreme values can shift the estimator heavily.

2.3. Sampling Based Approximation The first approximation, $T_1(A)$, provides us with a computationally efficient method for finding the epidemic threshold. This addresses the first issue pointed out at the beginning of Section 2.1. However, computing $T_1(A)$ requires data on the degree of all n

Table 2: Computational efficiency and statistical accuracy of $T_1(A)$

Model	Mean time $\lambda(A)$	Mean time $T_1(A)$	Mean error	SD error
Uniform	35.62 s	0.04 s	0.11%	0.03%
PowerLaw	33.45 s	0.04 s	3.66%	3.91%

nodes of the network. Therefore, this does not solve the second issue pointed out at the beginning of Section 2.1. We now propose a second alternative, T_2, to address the second issue. The idea behind this approximation is based on the same heuristic that was laid out in Section 2.2. Since $\lambda(P)$ is a function of degree moments, we can estimate these moments using observed node degrees. In defining $T_1(A)$, we used observed degrees of all n nodes in the network. However, we can also estimate the degree moments by considering a small sample of nodes, based on random walk sampling. The algorithm for computing T_2 is given in Algorithm 1.

Algorithm 1 RandomWalkEstimate.

procedure ESTIMATE(G, r, t^*)
  $x \leftarrow 1$.
  while $t \leq t^*$ **do**
    $x \leftarrow$ random neighbor of x, chosen uniformly.
  $v \leftarrow 0$.
  while $i \leq r$ **do**
    $v = v + d_x$
    $x \leftarrow$ random neighbor of x, chosen uniformly.
  return $T_2 = v/r$.

Note that we only use $(t^* + r)$ randomly sampled nodes for computing T_2, which implies that we do not need to collect or store data on the n individuals. Therefore this method overcomes the second issue pointed out at the beginning of Section 2.1. The approximation error arising from this method can be defined as

$$e_2(A) = \left| \frac{T_2(A)}{\lambda(A)} - 1 \right|, \tag{2.5}$$

and we want to show that $e_2(A) \to 0$ in probability, while the data-collection cost of $T_2(A)$ is much less than that of $T_1(A)$. In the next section, we are going to formalize this.

3 Theoretical Results on Approximation Errors

In this section, we are going to establish that the approximation errors $e_1(A)$ and $e_2(A)$, defined in Eqs. 2.4 and 2.5, converge to zero in probability. From Theorem 2.1 of Chung et al. (2003), we know that when

$$\frac{\sum_i \delta_i^2}{\sum_i \delta_i} > \log(n)\sqrt{\max_{1\le i\le n} \delta_i} \tag{3.1}$$

holds, then for any $\epsilon > 0$,

$$P\left[\left|\frac{\lambda(A)}{\lambda(P)} - 1\right| > \epsilon\right] \to 0.$$

Therefore, under Eq. 3.1, it suffices to show that, for any $\epsilon > 0$,

$$P\left[\left|\frac{T_1(A)}{\lambda(P)} - 1\right| > \epsilon\right] \to 0, \text{ and } P\left[\left|\frac{T_2(A)}{\lambda(P)} - 1\right| > \epsilon\right] \to 0.$$

To interpret the condition given in Eq. 3.1, suppose that the expected degrees are all of the same order, i.e., $\delta_i = O(n^\alpha)$ for some $\alpha \in (0, 1)$. Then, the left hand side of Eq. 3.1 is $O(n^\alpha)$, and the right hand side is $\log(n)O(n^{\alpha/2})$, which means the condition is satisfied for any $\alpha > 0$.

3.1. Convergence of $T_1(A)$ First, consider $T_1(A) = \frac{\sum_{i=1}^n d_i^2}{\sum_{i=1}^n d_i}$, and recall that $\lambda(P) = \frac{\sum_{i=1}^n \delta_i^2}{\sum_{i=1}^n \delta_i}$. For notational convenience, define $m_1 = \sum_{i=1}^n d_i, m_2 = \sum_{i=1}^n d_i^2, \mu_1 = \sum_{i=1}^n \delta_i, \mu_2 = \sum_{i=1}^n \delta_i^2$. We would like to show that, under reasonable conditions, for any $\epsilon > 0$,

$$P\left[\left|\frac{m_2\mu_1}{m_1\mu_2} - 1\right| > \epsilon\right] \to 0. \tag{3.2}$$

Next, we state the theorem which will establish a sufficient condition for this to hold. Please see Appendix for a proof of the theorem.

THEOREM 2. *If the average of the expected degrees goes to infinity, i.e., $\frac{1}{n}\sum_i \delta_i \to \infty$, and the spectral radius dominates $\log^2(n)$, i.e., $\frac{\sum_i \delta_i^2}{\sum_i \delta_i} = \omega(\log^2 n)$, then for any $\epsilon > 0$,*

$$P\left[\left|\frac{m_1}{\mu_1} - 1\right| > \epsilon\right] \to 0, \text{ and } P\left[\left|\frac{m_2}{\mu_2} - 1\right| > \epsilon\right] \to 0.$$

Thus, we have established that the approximation error for $T_1(A)$ goes to zero in probability. We have already observed in Section 2.2 that the runtime for $T_1(A)$ is orders of magnitude faster that the runtime for $\lambda(A)$. Therefore, $T_1(A)$ is both computationally efficient and statistically accurate as an approximation of the epidemic threshold.

3.2. Convergence of $T_2(A)$ Next, consider Algorithm 1. Let π denote the stationary distribution of the simple random walk on the given graph. Suppose the number of edges in the given graph is m. Recall that, π is given by $\pi_v = \frac{d_v}{\sum_v d_v}$ for all v. For brevity, we define the mixing time of the graph A, denoted as $t_{\text{mix}}(A)$, to mean the number of steps required by the simple random walk to reach a distribution $\hat{\pi}$ such that $\|\hat{\pi} - \pi\|_1 = o(\frac{1}{n^2})$. Let $T_2(A)$ be the estimate returned by the Algorithm 1. We first show an easy lemma that characterizes the bias of the estimator $T_2(A)$. Please see Appendix for a proof.

LEMMA 3. *If x is a node that is randomly sampled from π, and d_x is its degree, then $E[d_x] = \frac{\sum_i d_i^2}{\sum_i d_i}$. Consequently if $\hat{\pi}$ is such that $\|\pi - \hat{\pi}\|_1 = o(n^{-1})$ and x is sampled from $\hat{\pi}$, then $E[d_x] = (1 \pm o(1))\frac{\sum_i d_i^2}{\sum_i d_i}$.*

Next, we show that the estimator v_{RW} is actually concentrated around its expectation.

THEOREM 4 (Lezaud (1998)). *Let (X_n) be a irreducible and reversible Markov Chain on a finite set V with Q being the transition matrix. Let π be the stationary distribution. Let $f : V \to \Re$ be such that $E_\pi[f] = 0$, $\|f\|_\infty \leq 1$ and $0 < E_\pi[f^2] \leq b^2$. Then, for any initial distribution q, any positive integer r and all $0 < \gamma \leq 1$,*

$$\Pr_q\left[r^{-1}\sum_{i=1}^{r} f(X_i) \geq \gamma\right] \leq e^{-\varepsilon(Q)/5} S_q \exp\left(-\frac{r\gamma^2\varepsilon(Q)}{4b^2(1 + h(5\gamma/b^2))}\right),$$

where $\varepsilon(Q) = 1 - \lambda_2(Q)$, $\lambda_2(Q)$ being the second largest eigenvalue of Q, $S_q = \|q/\pi\|_2$ (in the $\ell_2(\pi)$ norm), and

$$h(x) = \frac{1}{2}(\sqrt{1+x} - (1 - x/2)).$$

If $\gamma \ll b^2$ and $\varepsilon(Q) \ll 1$, then the upper bound becomes

$$(1 + o(1))S_q \exp\left(-\frac{r\gamma^2\varepsilon(Q)}{4b^2(1 + o(1))}\right).$$

Using the above result, we bound the sample complexity of our estimator. We first quote the following result that we use to bound λ_1 of the transition matrix. Please see Appendix for a proof.

THEOREM 5. *Let $Q = D^{-1}A$. Let $\epsilon, \delta \in (0,1)$. Algorithm 1, using $r = \frac{1}{\varepsilon(Q)\epsilon^{3/2}} \times \frac{12md_{\max}}{(\sum_v d_v^2)} \log(1/\delta)$ and $t^* \geq t_{\text{mix}}(G)$ returns an estimate v_{RW} that satisfies, with probability $1-\delta$,*

$$(1-\epsilon)\frac{\sum_v d_v^2}{\sum_v d_v} \leq T_2(A) \leq (1+\epsilon)\frac{\sum_v d_v^2}{\sum_v d_v}.$$

The number of nodes that are touched by algorithm is $O(t^* + r)$.

Note that $Q = D^{-1}A$ has the same set of eigenvalues as the matrix $D^{-1/2}AD^{-1/2}$. For the Chung-Lu model, the eigenvalues of the matrix $L = I - D^{-1/2}AD^{-1/2}$ can be bounded by the following result from Chung et al. (2003).

THEOREM 6. *Let $L = I - D^{-1/2}AD^{-1/2}$ denote the normalized Laplacian. Let A be a random graph generated from the given expected degrees model, with expected degrees $\{\delta_i\}$, if the minimum expected degree $\delta_{\min}$ satisfies $\delta_{min} \gg \ln(n)$, then with probability at least $1 - 1/n = 1 - o(1)$, we have that for all eigenvalues $\lambda_k(L) > \lambda_{\min}(L)$ of the Laplacian of G,*

$$|1 - \lambda_k(L)| < 2\sqrt{\frac{6\ln(2n)}{\delta_{\min}}} = o(1).$$

It follows above that $\varepsilon(Q) = 1 - \lambda_2(Q) = 1 - \lambda_2(D^{-1/2}AD^{-1/2}) = \lambda_{n-1}(I - D^{-1/2}AD^{-1/2}) = 1 - o(1)$. Putting these together, we get the following corollary on the total number of node queries.

COROLLARY 6.1. *For a graph generated from the expected degrees model, with probability $1 - 1/n$, Algorithm 1, needs to query*

$$\ln(n) + \frac{1}{\epsilon^{3/2}} \times \frac{6(\sum_v d_v)d_{\max}}{(\sum_v d_v^2)} \log(1/\delta)$$

nodes in order to get a $(1 \pm \epsilon)$ estimate of $\sum_v d_v^2/2m$.

Note $\frac{6(\sum_v d_v)d_{\max}}{(\sum_v d_v^2)} \leq \frac{6d_{\max}}{d_{\min}}$, but this is a loose bound, better bounds can be derived for power law degree distributions, for instance.

Thus, we have proved that the approximation error for $T_2(A)$ goes to zero in probability. In addition, Corollary 6.1 shows that the number of nodes that we need to query in order to have an accurate approximation is much smaller than n. Furthermore, computing T_2 only requires node sampling and counting degrees, and therefore the runtime is much smaller than eigenvalue algorithms. Therefore, $T_2(A)$ is a computationally efficient and statistically accurate approximation of the epidemic threshold, while also requiring a much smaller data budget compared to $T_1(A)$.

4 Numerical Results

In this section, we characterize the empirical performance of our sampling algorithm on two synthetic networks, one generated from the Chung-Lu model and the second generated from the preferential attachment model of Barabási and Albert (1999).

4.1. Data Our first dataset is a graph generated from the Chung-Lu model of expected degrees. We generated a powerlaw sequence (i.e. fraction of nodes with degree d is proportion to $d^{-\beta}$) with exponent $\beta = 2.5$ and then generated a graph with this sequence as the expected degrees. Table 3 notes that, as expected, the first eigenvalue $\lambda_1(A)$ is close to $\frac{\sum_v d_v^2}{\sum_v d_v}$.

The second dataset is generated from the preferential attachment model (Barabási and Albert, 1999), where each incoming node adds 5 edges to the existing nodes, the probability of choosing a specific node as neighbor being proportional to the current degree of that node. While the preferential attachment model naturally gives rise to a directed graph, we convert the graph to an undirected one before running our algorithm. It is interesting to note that even in this case the Chung-Lu model does not hold, our first approximation, $T_1(A)$, is close to $\lambda(A)$.

4.2. Implementation Details In each of the networks, the random walk algorithm presented in Algorithm 1 was used for sampling. The random walk was started from an arbitrary node and every 10^{th} node was sampled (to account for the mixing time) from the walk. These samples were then used to calculate $T_2(A)$. This experiment was repeated 10 times. These gave estimates $T_2^1, \ldots, T_2^{10}$. We then calculate two relative errors $\forall i \in \{1, 2, \ldots, 10\}$,

$$\epsilon_i^{T_1-T_2} = \frac{\left|T_2^i - T_1(A)\right|}{T_1(A)},\ \epsilon_i^{\lambda-T2} = \frac{\left|T_2^i - \lambda(A)\right|}{\lambda(A)}.$$

We also note the following relation between the two error metrics.

$$\begin{aligned}\epsilon_i^{\lambda-T_2} &= \frac{|\lambda - T_2^i|}{\lambda} \leq \frac{|T_1 - T_2^i|}{\lambda} + \frac{|T_1 - T_2^i|}{\lambda} = \frac{|T_1 - T_2^i|}{\lambda} + \epsilon^{\lambda-T_1} = \frac{T_1}{\lambda}\epsilon_i^{T_1-T_2} + \epsilon^{\lambda-T_1} \\ &= (1 + \epsilon^{\lambda-T_1})\epsilon_i^{T_1-T_2} + \epsilon^{\lambda-T_1}.\end{aligned}$$

Table 3: Statistics of the two synthetic datasets used

Data	Nodes	Edges	$\lambda(A)$	$T_1(A)$	$\epsilon^{\lambda-T_1(A)}$
Chung-Lu ($\beta = 2.5$)	$50k$	$72k$	43.83	48.33	0.102
Chung-Lu (uniform)	$50k$	$130k$	67.60	67.46	0.002
Pref-Attach	$50k$	$250k$	37	32.8	0.128

We denote the averages of $\{\epsilon_i^{T_1-T_2}\}$ and $\{\epsilon_i^{\lambda-T_2}\}$ as $\epsilon^{T_1-T_2}$ and $\{\epsilon^{\lambda-T_2}\}$ respectively. It is easy to observe that the above relation holds between the two average quantities too.

We plot the averages $\epsilon^{T_1-T_2}$ and $\epsilon^{\lambda-T_2}$, along with the error-bars that reflect the standard deviation, against the *actual number of nodes seen by the random walk*. Note that the x-axis *accurately reflect how many times the algorithm actually queried the network*, not just the number of samples used. Measuring the cost of uniform node sampling in this setting, for instance, would need to keep track of how many nodes are touched by a Metropolis-Hastings walk that implements the uniform distribution.

4.3. Results In Fig. 1 We plot the two results for mean relative error, measure by $\epsilon_i^{\lambda-T_2}$ and $\epsilon_i^{T_1-T_2}$.

For the two Chung-Lu networks, the algorithm is able to get a 10% approximation to the statistic $T_1(A)$ by exploring at most 10% of the network. With more samples from the random walk, the mean relative errors settle to around 4–5%. However, once we measure the mean relative errors with respect to $\lambda(A)$, it becomes clearer that the estimator $T_2(A)$ does better when the graph is closer to the assumed (i.e. Chung-Lu) model. For the

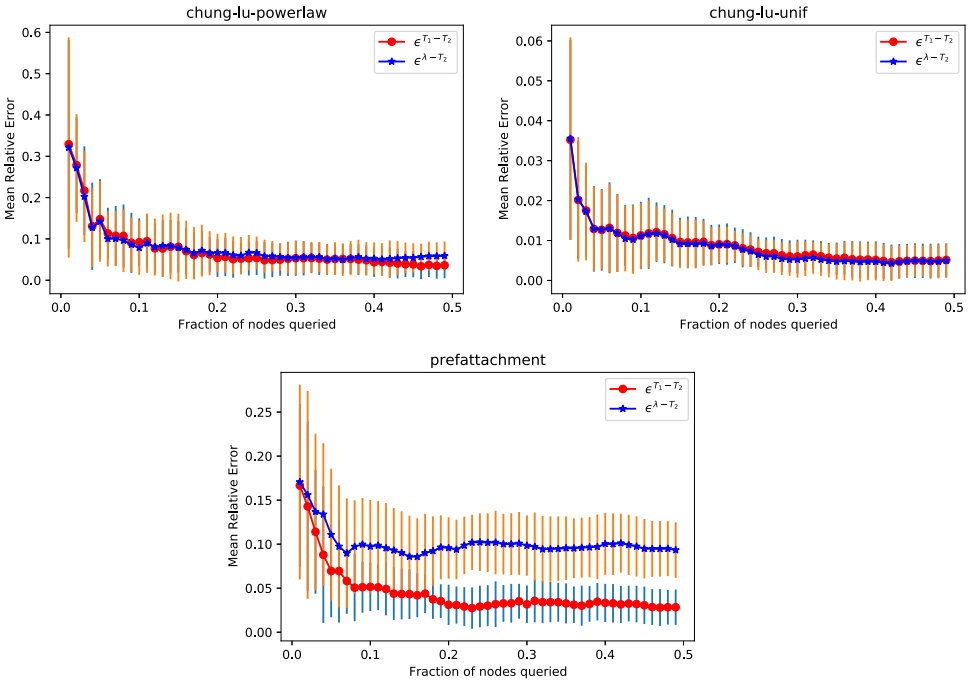

Figure 1: Results on three synthetic networks

Chung-Lu graph, the mean error $\epsilon^{\lambda-T2}$ essentially is very similar to $\epsilon^{T_1-T_2}$, which is to be expected. For the preferential attachment graph too, it is clear that the estimate T_2 is able to achieve a better than 10% relative error approximation of $\lambda(A)$.

Note that, if we were instead counting only the nodes whose degrees were actually used for estimation, the fraction of network used would be roughly 1–2% in all the cases, the majority of the node query cost actually goes in making the random walk mix, by using an initial burn-in period and by maintaining certain number of steps between subsequent samples.

5 Discussion

In this work, we investigated the problem of computing SIR epidemic thresholds of social contact networks from the perspective of statistical inference. We considered the two challenges that arise in this context, due to high computational and data-collection complexity of the spectral radius. For the Chung-Lu network generative model, the spectral radius can be characterized in terms of the degree moments. We utilized this fact to develop two approximations of the spectral radius. The first approximation is computationally efficient and statistically accurate, but requires data on observed degrees of all nodes. The second approximation retains the computationally efficiency and statistically accuracy of the first approximation, while also reducing the number of queries or the sample size quite substantially. The results seem very promising for networks arising from the Chung-Lu and preferential attachment generative models.

There are several interesting and important future directions. The methods proposed in this paper have provable guarantees only under the Chung-Lu model, although it works very well under the preferential attachment model. This seems to indicate that the degree based approximation might be applicable to a wider class of models. On the other hand, this leaves open the question of developing a better "model-free" estimator, as well as asking similar questions about other network features.

In this work we only considered the problem of accurate approximation of the epidemic threshold. From a statistical as well as a real-world perspective, there are several related inference questions. These include uncertainty quantification, confidence intervals, one-sample and two-sample testing, etc.

Social interaction patterns vary dynamically over time, and such network dynamics can have significant impacts on the contagion process Leitch et al. (2019). In this paper we only considered static social contact networks, and

in future we hope to study epidemic thresholds for time-varying or dynamic networks.

Finally, we note that the formulation in Eq. 2.1 is an approximation of the true epidemic threshold under the so-called quenched-mean-field approximation (Pastor-Satorras et al., 2015; Karrer et al., 2014). In recent work Castellano and Pastor-Satorras (2020), it has been shown that the SIS epidemic transition occurs at some point that is intermediate between $\lambda(A)$ and $T_1(A)$. In future work, we plan to extend our results to these more accurate expressions for the epidemic threshold.

Appendix A. Technical Proofs

A.1 Proof of Theorem 2

We will show that for any $\epsilon' > 0$,

$$P\left[\left|\frac{m_1}{\mu_1} - 1\right| > \epsilon'\right] \to 0, P\left[\left|\frac{m_2}{\mu_2} - 1\right| > \epsilon'\right] \to 0. \tag{A.1}$$

We first prove that Eq. A.1 implies Eq. 3.2. Equation A.1 implies that

$$P\left[\left\{\left|\frac{m_1}{\mu_1} - 1\right| > \epsilon'\right\} \cup \left\{\left|\frac{m_2}{\mu_2} - 1\right| > \epsilon'\right\}\right] \to 0.$$

Now, consider the event $\left\{\left|\frac{m_1}{\mu_1} - 1\right| \leq \epsilon'\right\} \cap \left\{\left|\frac{m_2}{\mu_2} - 1\right| \leq \epsilon'\right\}$. Note that m_2/m_1 is a strictly increasing function of m_2 and a strictly decreasing function of m_1. Therefore, for outcomes belonging to the above event,

$$\frac{\mu_2}{\mu_1} \times \frac{1-\epsilon'}{1+\epsilon'} \leq \frac{m_2}{m_1} \leq \frac{\mu_2}{\mu_1} \times \frac{1+\epsilon'}{1-\epsilon'}.$$

Note that

$$1 - \frac{1-\epsilon'}{1+\epsilon'} = \frac{2\epsilon'}{1+\epsilon'} < 2\epsilon', \text{ and } \frac{1+\epsilon'}{1-\epsilon'} - 1 = \frac{2\epsilon'}{1-\epsilon'} < 4\epsilon',$$

given that $\epsilon' < 1/2$. Now, fix $\epsilon > 0$ and let $\epsilon' = \epsilon/4$. Then,

$$Eq. \quad \Rightarrow P\left[\left|\frac{m_2\mu_1}{m_1\mu_2} - 1\right| > 4\epsilon'\right] \to 0 \Rightarrow Eq. \quad .$$

Thus, proving Eq. A.1 is sufficient for proving Eq. 3.2.

Proof of Theorem 2. We will use Hoeffding's inequality (Hoeffding, 1994) for the first part, and we begin by stating the inequality for the sum

of Bernoulli random variables. Let $B_1, \ldots, B_m$ be m independent (but not necessarily identically distributed) Bernoulli random variables, and $S_m = \sum_{i=1}^{m} B_i$. Then for any $t > 0$,

$$P[|S_m - E[S_m]| \geq t] \leq 2 \exp\left(\frac{-2t^2}{m}\right).$$

In our case,

$$m_1 = \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} \sum_{j=1}^{n} A(i,j) = 2 \sum_{i<j} A(i,j),$$

and we know that $\{A(i,j) : 1 \leq i < j \leq n\}$ are independent Bernoulli random variables. Fix $\epsilon > 0$ and note that $E[\sum_{i<j} A(i,j)] = \frac{1}{2}\mu_1$. Using Hoeffding's inequality with $S_m = m_1/2$, $m = \binom{n}{2}$, and $t = \frac{\epsilon}{2}\mu_1$, we get

$$P\left[\left|\frac{m_1}{2} - \frac{\mu_1}{2}\right| > \frac{\epsilon}{2}\mu_1\right] \leq 2 \exp\left(-\epsilon^2 \frac{\mu_1^2}{n(n-1)}\right).$$

Since $\frac{1}{n}\sum_i \delta_i \to \infty$, the right hand side goes to zero. Therefore,

$$P\left[\left|\frac{m_1}{\mu_1} - 1\right| > \epsilon\right] \to 0.$$

For the second part, we can characterize m_2 as following.

$$E[m_2] = E\left[\sum_i d_i^2\right] = \sum_i (E[d_i])^2 + var(d_i) = \mu_2 + var(d_i),$$

and hence,

$$|m_2 - \mu_2| \leq |m_2 - E[m_2]| + |E[m_2] - \mu_2|.$$

We show that, under the given assumptions, with probability $1 - o(1)$, $|m_2 - E[m_2]| = o(\mu_2)$. Furthermore, $|E[m_2] - \mu_2| = o(\mu_2)$.

As noted before, each d_i is a sum of binomial random variables. By applying Chernoff-Hoeffding bound, and union bounding over all $i \in \{1, \ldots, n\}$, we can get, with probability $1 - o(1)$, and for any fixed $\epsilon \in (0,1)$,

$$\forall i \in \{1, \ldots, n\},\ d_i \leq \delta_i + \max\{\epsilon\delta_i, O(\log(n))\}.$$

Let the above event be called the event $\mathcal{A}$. If the event $\mathcal{A}$ happens, then,

$$m_2 = \sum_i d_i^2 \quad \leq \quad \sum_i \delta_i^2 + 2\delta_i \max(\epsilon\delta_i, O(\log(n))) + \max(\epsilon^2\delta_i^2, O(\log^2 n))$$

$$\begin{aligned} &\leq \mu_2 + 2\sum_i \delta_i(\epsilon\delta_i + O(\log(n))) + (\epsilon^2\delta_i^2 + O(\log^2 n)) \\ &\leq \mu_2 + 3\epsilon\mu_2 + (n + \sum_i \delta_i)O(\log^2 n) \\ \left|\frac{m_2}{\mu_2} - 1\right| &\leq 3\epsilon + (n + \sum_i \delta_i)O(\log^2 n)/\mu_2. \end{aligned}$$

Note that $\frac{n}{\mu_2} = \frac{1}{\sum_i \delta_i^2/n} \to 0$ under the given assumption. Furthermore,

$$\frac{(\sum_i \delta_i)O(\log^2 n)}{\sum_i \delta_i^2} = o(1) \to 0.$$

Putting these together, and using $\epsilon' = 3\epsilon$ we have the given claim.

A.2 Proof of Theorem 5

Proof of Lemma 3. It is easy to see that

$$E_{x\sim\pi}[d_x] = \sum_{v=1}^{n} d_v \times \pi_v = \frac{\sum_v d_v^2}{\sum_v d_v}.$$

We show the second claim as follows:

$$|E_{x\sim\pi}[d_x] - E_{x\sim\hat{\pi}}[d_x]| \leq \sum_{v=1}^{n} d_v|\pi_v - \hat{\pi_v}| \leq n\|\pi - \hat{\pi}\|_1 = o(1).$$

Proof of Theorem 5. In our setting the set V is the set of vertices. Define the function $f(X_i)$ as :

$$d_{\max} \times f(X_i) = d_{X_i} - E_\pi[d_{X_i}].$$

$f(\cdot)$ clearly satisfies $E_\pi[f] = 0$ and that $\|f\|_\infty \leq 1$. We can bound $E_\pi[f^2]$ as

$$E_\pi[f^2] \leq d_{\max}^{-2} E_\pi[d_v^2] = d_{\max}^{-2} \sum_v \frac{d_v^2 \times d_v}{\sum_v d_v} = d_{\max}^{-2} \sum_v \frac{d_v^3}{\sum_v d_v}.$$

Using the first t^* steps, we reach the distribution $\hat{\pi}$ that satisfies $\|\pi - \hat{\pi}\|_1 = o(n^{-1})$. Hence,

$$\|\hat{\pi}/\pi\|_2^2 = \sum_v \pi_v(\hat{\pi}_v/\pi_v)^2 = \sum_v \hat{\pi}_v^2/\pi_v = \sum_v (\pi_v + (\hat{\pi}_v - \pi_v))^2/\pi_v$$

$$\begin{aligned} &= \sum_v (\pi_v + 2(\hat{\pi}_v - \pi_v) + (\hat{\pi}_v - \pi_v)^2/\pi_v) \\ &= 1 + 2 \times (1-1) + \sum_v (\hat{\pi}_v - \pi_v)^2/\pi_v \leq 1 + \|\pi - \hat{\pi}\|_2^2 / \min(\pi_v) \\ &\leq 1 + \|\pi - \hat{\pi}\|_1^2 \left(\sum_v d_v \right) / d_{\min} = 1 + o(1), \end{aligned}$$

where the last step follows as $\|\pi - \hat{\pi}\|_1 = o(n^{-2})$.

We use $b^2 = d_{\max}^{-2} \sum_v \frac{d_v^3}{\sum_v d_v}$ and $\gamma = \epsilon d_{\max}^{-1} \times \frac{\sum_v d_v^2}{\sum_v d_v}$. Hence

$$\gamma/b^2 = \epsilon d_{\max} \frac{\sum_v d_v^2}{\sum_v d_v^3} \text{ and } \gamma^2/b^2 = \epsilon^2 \frac{(\sum_v d_v^2)^2}{(\sum_v d_v)(\sum_v d_v^3)}.$$

Hence,

$$\begin{aligned} h(5\gamma/b^2) &= \left(1 + 5\epsilon d_{\max} \frac{\sum_v d_v^2}{\sum_v d_v^3}\right)^{1/2} - 1 + 5\epsilon d_{\max} \frac{\sum_v d_v^2}{2\sum_v d_v^3} \\ &\leq \left(5\epsilon d_{\max} \frac{\sum_v d_v^2}{\sum_v d_v^3}\right)^{1/2} + 2.5\epsilon d_{\max} \frac{\sum_v d_v^2}{\sum_v d_v^3} \\ &\leq 6\epsilon^{1/2} d_{\max} \frac{\sum_v d_v^2}{\sum_v d_v^3}. \end{aligned}$$

Plugging this, we get that

$$\begin{aligned} \frac{r\gamma^2 \varepsilon(Q)}{4b^2(1 + h(5\gamma/b^2))} &\geq r\varepsilon(Q) \times \epsilon^2 \frac{(\sum_v d_v^2)^2}{(\sum_v d_v)(\sum_v d_v^3)} \times \left(1 + 6\epsilon^{1/2} d_{\max} \frac{\sum_v d_v^2}{\sum_v d_v^3}\right)^{-1} \\ &\geq \frac{r\varepsilon(Q)\epsilon^{3/2}(\sum_v d_v^2)}{6(\sum_v d_v) d_{\max}}. \end{aligned}$$

Setting $r = \frac{1}{\varepsilon(Q)\epsilon^{3/2}} \times \frac{6(\sum_v d_v) d_{\max}}{(\sum_v d_v^2)} \log(1/\delta)$, and using Theorem 4, we can claim that, with probability $1 - \delta$,

$$T_2(A) \in \left((1-\epsilon) \frac{\sum_v d_v^2}{\sum_v d_v}, (1+\epsilon) \frac{\sum_v d_v^2}{\sum_v d_v} \right).$$

The bound on the number of nodes touched/queried by the algorithm follows naturally.

Disclaimer with Respect to Current Pandemic

We do realize that in the face of the current pandemic, while it is important to pursue research relevant to it, it is also important to be responsible in following the proper scientific process. We would like to state that in this work, the question of epidemic threshold estimation has been formalized from a theoretical viewpoint in a much used, but simple, random graph model. We are not yet at a position to give any guarantees about the performance of our estimator in real social networks. We do hope, however, that the techniques developed here can be further refined to work to give reliable estimators in practical settings.

Acknowledgements. We thank the Associate Editor and two anonymous reviewers for their constructive suggestions, which were really helpful towards the improvement of the manuscript. Anirban acknowledges the kind support of the N. Rama Rao Chair Professorship at IIT Gandhinagar, the Google India AI/ML award (2020), Google Faculty Award (2015), and CISCO University Research Grant (2016). Srijan acknowledges the support from an NIH R01 grant 1R01LM013309.

References

AIELLO, W., CHUNG, F. and LU, L. (2000). A random graph model for massive graphs, In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of computing.* ACM, p. 171–180.

BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509–512.

BARRETT, C.L., BISSET, K.R., EUBANK, S.G., FENG, X. and MARATHE, M.V. (2008). Episimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks, In *SC'08: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing.* IEEE, p. 1–12.

BENAYCH-GEORGES, F., BORDENAVE, C., KNOWLES, A. et al. (2019). Largest eigenvalues of sparse inhomogeneous erdős–rényi graphs. *Ann. Probab.* **47**, 1653–1676.

BENGTSSON, L., GAUDART, J., LU, X., MOORE, S., WETTER, E., SALLAH, K., REBAUDET, S. and PIARROUX, R. (2015). Using mobile phone data to predict the spatial spread of cholera. *Sci. Rep.* **5**, 8923.

BHADRA, S., CHAKRABORTY, K., SENGUPTA, S. and LAHIRI, S. (2019). A bootstrap-based inference framework for testing similarity of paired networks. arXiv:1911.06869.

BICKEL, P.J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci.* **106**, 21068–21073.

BICKEL, P.J. and SARKAR, P. (2016). Hypothesis testing for automated community detection in networks. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **78**, 253–273.

BORDENAVE, C., BENAYCH-GEORGES, F. and KNOWLES, A (2020). Spectral radii of sparse random matrices. *Ann. l'Inst. Henri Poincare (B) Probab. Stat.*

BRAUER, F. and CASTILLO-CHAVEZ, C. (2012). *Mathematical models in population biology and epidemiology, vol. 2.* Springer, Berlin.

CASTELLANO, C. and PASTOR-SATORRAS, R. (2020). Cumulative merging percolation and the epidemic transition of the susceptible-infected-susceptible model in networks. *Phys. Rev. X* **10**, 011070.

CHAKRABARTI, D., WANG, Y., WANG, C., LESKOVEC, J. and FALOUTSOS, C. (2008). Epidemic thresholds in real networks. *ACM Trans. Inf. Syst. Secur.* **10**, 1–26.

CHINAZZI, M., DAVIS, J.T., AJELLI, M., GIOANNINI, C., LITVINOVA, M., MERLER, S., PIONTTI, A.P., MU, K., ROSSI, L., SUN, K. and ET AL. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science* **368**, 6489, 395–400.

CHUNG, F. and LU, L. (2002). The average distances in random graphs with given expected degrees. *Proc. Natl. Acad. Sci.* **99**, 15879–15882.

CHUNG, F. and RADCLIFFE, M. (2011). On the spectra of general random graphs. *Electron. J. Combinator.* **18**, P215–P215.

CHUNG, F., LU, L. and VU, V. (2003). Eigenvalues of random power law graphs. *Ann. Combinator.* **7**, 21–33.

COLIZZA, V. and VESPIGNANI, A. (2007). Invasion threshold in heterogeneous metapopulation networks. *Phys. Rev. Lett.* **99**, 148701.

DALLAS, T.A., KRKOŠEK, M. and DRAKE, J.M. (2018). Experimental evidence of a pathogen invasion threshold. *R. Soc. Open Sci.* **5**, 171975.

DECREUSEFOND, L., DHERSIN, J. -S., MOYAL, P., TRAN, V.C. et al. (2012). Large graph limit for an sir process in random network with heterogeneous connectivity. *Ann. Appl. Probab.* **22**, 541–575.

EUBANK, S., GUCLU, H., KUMAR, V.A., MARATHE, M.V., SRINIVASAN, A., TOROCZKAI, Z. and WANG, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature* **429**, 180–184.

GALVANI, A.P. and MAY, R.M. (2005). Dimensions of superspreading. *Nature* **438**, 293–295.

GHOSHDASTIDAR, D. and VON LUXBURG, U. (2018). Practical methods for graph two-sample testing, In *Advances in Neural Information Processing Systems*, p. 3019–3028.

GÓMEZ, S., ARENAS, A., BORGE-HOLTHOEFER, J., MELONI, S. and MORENO, Y. (2010). Discrete-time markov chain approach to contact-based disease spreading in complex networks. *EPL (Europhys. Lett.)* **89**, 38009.

HANDCOCK, M.S., RAFTERY, A.E. and TANTRUM, J.M. (2007). Model-based clustering for social networks. *J. R. Stat. Soc.: Ser. A* **170**, 301–354.

HETHCOTE, H.W. (2000). The mathematics of infectious diseases. *SIAM Rev.* **42**, 599–653.

HOEFFDING, W. (1994). Probability inequalities for sums of bounded random variables, In *The Collected Works of Wassily Hoeffding*. Springer, p. 409–426.

HOFF, P.D., RAFTERY, A.E. and HANDCOCK, M.S. (2002). Latent space approaches to social network analysis. *J. Am. Stat. Assoc.* **97**, 1090–1098.

HUANG, C., WANG, Y., LI, X., REN, L., ZHAO, J., HU, Y., ZHANG, L., FAN, G., XU, J., GU, X. and ET AL. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506.

KARRER, B., NEWMAN, M.E. and ZDEBOROVÁ, L. (2014). Percolation on sparse networks. *Phys. Rev. Lett.* **113**, 20, 208702.

KEELING, M. (2005). The implications of network structure for epidemic dynamics. *Theor. Popul. Biol.* **67**, 1–8.

KERMACK, W.O. and MCKENDRICK, A.G. (1927). A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A, Containing papers of a mathematical and physical character* **115**, 700–721.

KERMACK, W.O. and MCKENDRICK, A.G. (1932). Contributions to the mathematical theory of epidemics. ii.—the problem of endemicity. *Proc. R. Soc. Lond. Ser. A, Containing papers of a mathematical and physical character* **138**, 55–83.

KERMACK, W.O. and MCKENDRICK, A.G. (1933). Contributions to the mathematical theory of epidemics. iii.—further studies of the problem of endemicity. *Proc. R. Soc. Lond. Ser. A, Containing Papers of a Mathematical and Physical Character* **141**, 94–122.

KOMOLAFE, T., QUEVEDO, A.V., SENGUPTA, S. and WOODALL, W.H. (2019). Statistical evaluation of spectral methods for anomaly detection in static networks. *Netw. Sci.* **7**, 319–352.

KRAMER, A.M., PULLIAM, J.T., ALEXANDER, L.W., PARK, A.W., ROHANI, P. and DRAKE, J.M. (2016). Spatial spread of the west africa ebola epidemic. *R. Soc. Open Sci.* **3**, 8, 160294.

KRIVITSKY, P.N., HANDCOCK, M.S., RAFTERY, A.E. and HOFF, P.D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Netw.* **31**, 204–213.

LEITCH, J., ALEXANDER, K.A. and SENGUPTA, S. (2019). Toward epidemic thresholds on temporal networks: a review and open questions. *Appl. Netw. Sci.* **4**, 105.

LEZAUD, P. (1998). Chernoff-type bound for finite markov chains. *Ann. Appl. Probab.* **8**, 3, 849–867.

MEYERS, L.A., POURBOHLOUL, B., NEWMAN, M., SKOWRONSKI, D.M. and BRUNHAM, R.C. (2005). Network theory and SARS: predicting outbreak diversity. *J. Theor. Biol.* **232**, 71–81.

NEWMAN, M.E.J. (2002). Spread of epidemic disease on networks. *Phys. Rev. E* **66**, 1, 016128.

PASTOR-SATORRAS, R., CASTELLANO, C., VAN MIEGHEM, P. and VESPIGNANI, A. (2015). Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925–979.

PINAR, A., SESHADHRI, C. and KOLDA, T.G. (2012). The similarity between stochastic Kronecker and Chung-lu graph models, In *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, p. 1071–1082.

POURBOHLOUL, B., MEYERS, L., SKOWRONSKI, D., KRAJDEN, M., PATRICK, D. and BRUNHAM, R. (2005). Modeling control strategies of respiratory pathogens. *Emerg. Infect. Dis.* **11**, 1249–56.

PRAKASH, B.A., CHAKRABARTI, D., FALOUTSOS, M., VALLER, N. and FALOUTSOS, C. (2010). Got the flu (or mumps)? Check the Eigenvalue! arXiv:1004.0060.

ROCHA, L.E.C., LILJEROS, F. and HOLME, P. (2011). Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Comput. Biol.* **7**, e1001109.

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Stat.* **39**, 1878–1915.

SENGUPTA, S. (2018). Anomaly detection in static networks using egonets. arXiv:1807.08925.

SENGUPTA, S. and CHEN, Y. (2015). Spectral clustering in heterogeneous networks. *Stat. Sin.* **25**, 1081–1106.

SENGUPTA, S. and CHEN, Y. (2018). A block model for node popularity in networks with community structure. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **80**, 365–386.

SHULGIN, B., STONE, L. and AGUR, Z. (1998). Pulse vaccination strategy in the sir epidemic model. *Bull. Math. Biol.* **60**, 1123–1148.

SUN, K., CHEN, J. and VIBOUD, C. (2020). Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. *Lancet Digit. Health* **2**, 4, e201–e208.

Tang, M., Athreya, A., Sussman, D.L., Lyzinski, V., Park, Y. and Priebe, C.E. (2017a). A semiparametric two-sample hypothesis testing problem for random graphs. *J. Comput. Graph. Stat.* **26**, 344–354.

Tang, M., Athreya, A., Sussman, D.L., Lyzinski, V. and Priebe, C.E. (2017b). A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli* **23**, 1599–1630.

van den Driessche, P. and Watmough, J. (2002). Reproduction numbers and subthreshold endemic equilibria for compartmental models of disease transmission. *Math. Biosci.* **180**, 29–48.

Wallinga, J., Heijne, J.C. and Kretzschmar, M. (2005). A measles epidemic threshold in a highly vaccinated population. *PLoS Med.* **2**, e316.

Wang, Y.R. and Bickel, P.J. (2017). Likelihood-based model selection for stochastic block models. *Ann. Stat.* **45**, 500–528.

Wang, Y., Chakrabarti, D., Wang, C. and Faloutsos, C. (2003). Epidemic spreading in real networks: an eigenvalue viewpoint, In *22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings.* IEEE Computer Society, Florence, p. 25–34.

Wang, W., Liu, Q.H., Zhong, L.F. and et al. (2016). Predicting the epidemic threshold of the susceptible-infected-recovered model. *Sci. Rep.* **6**, 24676. doi: 10.1038/srep24676.

Wang, W., Tang, M., Stanley, H.E. and Braunstein, L.A. (2017). Unification of theoretical approaches for epidemic spreading on complex networks. *Rep. Progr. Phys.* **80**, 036603.

Wang, C., Horby, P.W., Hayden, F.G. and Gao, G.F. (2020). A novel coronavirus outbreak of global health concern. *Lancet* **395**, 470–473.

Woolhouse, M.E.J., Dye, C., Etard, J.F., Smith, T., Charlwood, J.D., Garnett, G.P., Hagan, P., Hii, J.L.K., Ndhlovu, P.D., Quinnell, R.J., Watts, C.H., Chandiwana, S.K. and Anderson, R.M. (1997). Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proc. Natl. Acad. Sci.* **94**, 338–342.

Yan, X., Shalizi, C., Jensen, J.E., Krzakala, F., Moore, C., Zdeborová, L., Zhang, P. and Zhu, Y. (2014). Model selection for degree-corrected block models. *J. Stat. Mech.: Theory Exp.* **2014**, P05007.

Zhang, X., Moore, C. and Newman, M.E. (2017). Random graph models for dynamic networks. *Eur. Phys. J. B* **90**, 200.

Zhao, Y., Levina, E. and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Stat.* **40**, 2266–2292.

Zhao, M.J., Driscoll, A.R., Sengupta, S., Fricker, Jr. R.D., Spitzner, D.J. and Woodall, W.H. (2018). Performance evaluation of social network anomaly detection using a moving window–based scan method. *Qual. Reliab. Eng. Int.* **34**, 1699–1716.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R. and et al. (2020). A novel coronavirus from patients with pneumonia in China. *New Engl. J. Med.*, 2019.

Anirban Dasgupta
Computer Science and Engineering,
Indian Institute of Technology,
Gandhinagar, Gandhinagar, India
E-mail: anirbandg@iitgn.ac.in

Srijan Sengupta
Statistics, North Carolina State
University, Raleigh, NC, USA
E-mail: sengupta@vt.edu

Paper received: 16 March 2020; accepted 11 May 2021.