Epidemic Forecasting on Networks: Bridging Local Samples with Global Outcomes

Yeganeh Alimohammadi
University of California Berkeley, yeganeha@berkeley.edu

Christian Borgs
University of California Berkeley, borgs@berkeley.edu

Remco van der Hofstad
Eindhoven University of Technology, r.w.v.d.hofstad@tue.nl

Amin Saberi
Stanford University, saberi@stanford.edu

We study Susceptible-Infected (SI), Susceptible-Infected-Removed (SIR), and related epidemic models in which infected individuals transition to an absorbing state, such as recovery or permanent infectiousness. In addition to infectious diseases, these models are used for studying the diffusion of innovations in which new behaviors, opinions, conventions, and technologies propagate from person to person through a social network.

We focus on the key challenge of forecasting epidemic trajectory and outbreak sizes and show that they can be predicted with a few samples from the network data. To this end, we propose a local algorithm for epidemic estimation, and prove the estimator's accuracy for both deterministic finite graphs and random networks, given certain neighborhood constraints. Further, leveraging the theory of local graph limits, we relate the time evolution in a sequence of graphs converging locally in probability with the epidemic in the limit graph. Finally, we validate our findings with experiments on synthetic models and real-world networks, such as Copenhagen and San Francisco's SafeGraph data.

Key words: SIR epidemics, local convergence, local estimation

1. Introduction

Epidemic models, originally conceived by Bernoulli (1760) and gaining prominence during the Spanish flu epidemics in the early 20th century (Ross and Hudson 1917, Kermack and McKendrick 1927), are developed for and frequently applied to the analysis of the spread of infectious diseases. These models categorize populations into various compartments such as Susceptible (S), Infectious

(I), or Recovered (R), representing potential flow patterns individuals may experience throughout the course of an epidemic. As instrumental tools in public health policy formulation, they assist in forecasting various aspects of an epidemic, including its spread, the total number of infections, and its duration (Eubank et al. 2004, Larson 2007, Lloyd-Smith et al. 2009, Heesterbeek et al. 2015, Scarpino and Petri 2019). Furthermore, they aid in evaluating the potential impacts of health interventions, serving as a guide in optimizing strategies such as the allocation of limited vaccine resources or targeted closures, thus playing a central role in assessing the efficacy and selection of countermeasures during public health emergencies (Wu et al. 2005, Mamani et al. 2013, Kaplan 2020, Bastani et al. 2021, Birge et al. 2022, Acemoglu et al. 2023).

Beyond the scope of infectious diseases, epidemic models also provide a framework for understanding the diffusion of innovations in social networks where new behaviors, conventions, and technologies spread from person to person (Bass 1969). Such analyses offer insight into “word-of-mouth” effects, and have informed our understanding of medical and agricultural innovations, viral marketing, cascading failures in power systems, and the spread of misinformation (Kalish and Lilien 1983, Ford et al. 2006, Feder and Umali 1993, Lee et al. 2015, Amini and Minca 2016, Yang et al. 2017, Mostagir and Siderius 2023). Recent research further explores the fundamental algorithmic problems in these systems, leveraging network data to optimize marketing strategies targeting influential network members, thereby enhancing the adoption rate of new products (Kempe et al. 2003, Goel et al. 2016, Lobel et al. 2017, Ajorlou et al. 2018, Akbarpour et al. 2018, Manshadi et al. 2020, Chin et al. 2022).

Predicting the trajectory of the epidemic is a central challenge across the mentioned domains. Traditional approaches of forecasting epidemic trajectories tend to fall into two categories: On one hand, mean-field and random network models offer a model-based perspective that relies on simplifying assumptions such as uniform mixing or a specific network topology (Bartlett 1949, Britton et al. 2019, Dimitrov and Meyers 2010, Mukherjee and Seshadri 2022, Bampo et al. 2008, Manshadi et al. 2020, Kiss et al. 2017). These methods provide analytical insight but do not directly incorporate real-world data. On the other hand, full-scale simulation methods use granular data such as mobile tracking to trace epidemic progression in detail (Bajardi et al. 2011, Wesolowski et al. 2012, Chang et al. 2021). However, in many practical situations, such complete data are simply not available.

In many real-world situations, data are typically obtained via methods like contact tracing, which capture only a small, local portion of the network. Motivated by this, we propose a distinct, data-driven approach that relies on the collection of small samples of local network data to predict an epidemic. Recent work has ventured into similar methodologies, particularly in seeding strategies and estimating the final size of an outbreak, (Eckles et al. 2022, Alimohammadi et al. 2022). However, their scope differs from what we present here. For an in-depth comparison, see Section 1.2.

Our proposed algorithm uses samples from the interaction networks to approximate the future trajectory of an epidemic. It leverages the local neighborhood of randomly chosen individuals to create an estimator for the proportion of the population in each state of the epidemic at any given time. We derive bounds on the estimator's error on any fixed graph and show that, under certain assumptions, the required sample size to achieve a prediction with ε additive error is independent of the network size. Our findings, applicable to both deterministic graphs and a broad class of random network models, suggest that accessing the local network structure of a few individuals is sufficient to estimate the time evolution of the epidemic.

1.1. Summary of Our Contribution

Our work makes several key contributions:

1) Algorithmic Estimation: We introduce a *local algorithm* for estimating the epidemic's time evolution by simulating the epidemic process in the local neighborhood of a few random nodes going backward in time. Using the Susceptible–Infectious–Recovered (SIR) model as an illustrative example, our algorithm assumes that the contact-time distribution D_I and the recovery time distribution D_R are known (or can be estimated from prior data with some error). Under the SIR model, an infected node recovers after a time drawn from D_R and, while infectious, transmits the disease at a rate governed by D_I. Given an initial node v and a probing budget k, the algorithm simulates v's status (susceptible, infectious, or recovered) at any time $t \geq 0$, by performing a backward search from v that probes at most k nodes in its local neighborhood. By averaging the outcomes of this algorithm with independent starting points, we construct an estimator for the time evolution of epidemics. For details, see Section 2.1.

2) Theoretical Guarantees of the Estimator: We give bounds on the error of our estimators for any given deterministic graph (Theorem 1). Under the assumption of moderate growth in local subgraph sizes (formalized by a tightness condition), we prove that the required sample size for an ε additive error is a constant independent of the network size (Theorem 7). We then extend our result to random network models, where we bound the error of our estimator under a condition that ensures the empirical distributions of local neighborhood structures remain stable across different network realizations (Theorem 2). Similar to the deterministic case, the required sample size depends solely on the target prediction accuracy, suggesting that the time evolution of epidemics can be predicted by probing the local network structure of just a few individuals.

Moreover, we generalize our results to allow *weighted* sampling—where nodes are chosen with probabilities reflecting local attributes (e.g., community membership). We also observe a natural tradeoff: increasing the probing budget k reduces bias (by better approximating the global epidemic dynamics), but increases variance (due to overlapping neighborhoods); see Section 2.3. Finally, while our theoretical analysis assumes known epidemic parameters D_I and D_R, these can be estimated from data, and our framework is robust to moderate estimation errors (Section 2.4).

3) Asymptotic Characterization: We also study the asymptotics of epidemics on a sequence of graphs of growing sizes. Leveraging the theory of local graph limits (Benjamini and Schramm 2001, Aldous and Steele 2004), we prove that the epidemic's time evolution in a sequence of graphs with a local limit in probability converges to the same process in the graph limit (Theorem 3). This implies that the time evolution of epidemics is essentially a 'local' property of the graph.

4) Applicability to General Epidemic Models: Our framework extends naturally to more complex epidemic models that incorporate intermediate states or time-varying infectiousness. For example, our model can accommodate scenarios where a susceptible individual becomes exposed, transitions through stages of varying transmission rates, and ultimately recovers. Moreover, the initial infection configuration may depend on local network attributes such as degree (see Section 2.5).

In the context of real-world applications, our framework encompasses several well-established epidemic models, including those with heterogeneous infectiousness for HIV (May and Anderson 1987, Isham 1988), malaria (Mandal et al. 2011, Gupta et al. 1994), models with carrier states for tuberculosis (Aparicio et al. 2000, Blower et al. 1995) and typhoid (Cvjetanović et al. 1971), influenza (Andreasen et al. 1997), COVID-19 (Bertsimas et al. 2021, Mukherjee and Seshadri 2022), and even models of information cascade (Watts 2002) and viral marketing models (Bass 1969, Jackson and Yariv 2005, Banerjee et al. 2013, Bampo et al. 2008, Ajorlou et al. 2018, Manshadi et al. 2020). A central theme behind these models is that they all reach an eventual absorbing state, whether infectiousness or recovery. Thus, it is possible to run the epidemic process backward and apply our algorithm.

5) Experimental Validation: We empirically validate our theoretical finding by conducting experiments on synthetic and real-world networks, including the Copenhagen Interaction Network created by the Bluetooth data of over 400 students (Sapiezynski et al. 2019) and San Francisco SafeGraph data with more than 30,000 nodes representing census blocks and points of interest across San Francisco (Chang et al. 2021). Notably, the San Francisco SafeGraph dataset includes edge weights, representing the transmission strength between nodes. In our experiments, we compare the outcomes of running an SIR epidemic on the entire graph against our estimator's results, which uses local network information of a few nodes. Remarkably, even with access to less than 1% of the nodes in the San Francisco dataset and approximately 14% of nodes in the Copenhagen dataset, our methodology yields predictions that align closely with the actual epidemic trajectories, falling within the 95% confidence interval of the true time evolution (see details in Section 5).

1.2. Related Work

In the intersection of epidemic prediction and operations research, a recent survey by (Gupta et al. 2022) highlights the need for refined models and efficient data collection mechanisms. We address this gap, by leveraging small samples of network data to improve epidemic predictions, contributing to two main research avenues: understanding epidemics with small data, and epidemics asymptotics.

Prediction with small data: Harnessing the power of small data in epidemic modeling has gained attention in the past few years. Recent work by Baek et al. (2021) studied sample complexity for estimating diffusion models, including SIR, with small data and unobservable networks, deriving lower bounds on the number of samples required to estimate outbreak size—bounds that increase with population size when network data are absent. This underscores the advantage of incorporating network structure, as our method does, to reduce sample size.

Recognizing the need for network data, several studies showcase its pivotal role in enhancing predictions, especially regarding influence maximization. These studies typically lean on heuristics for probing the network data (Mihara et al. 2015, Stein et al. 2017, Chen et al. 2022) or offer theoretical perspectives tailored to specific random graph models (Wilder et al. 2018).

In this vein, two studies (Eckles et al. 2022, Alimohammadi et al. 2022) investigate small network data under the independent cascade model, a variant of the SIR model where nodes transmit disease or information with fixed probability p. The primary goal of (Alimohammadi et al. 2022) is to approximate the epidemic's final size, closely related to the size of the largest component under percolation — a mathematical structure where each edge of the graph is retained with a given probability p. They propose a local algorithm for this purpose. Their algorithm, bearing some resemblance to ours without a temporal element, starts from a random node and outputs an estimate of the number of nodes it can infect. Under the assumption that the graph is an expander, they prove that constant queries to this algorithm yield an $(1-\epsilon)$ approximation of the infection's end size, using their earlier work on percolation (Alimohammadi et al. 2023). Contrasting with their work, our focus extends beyond the final infection size to include the epidemic's time evolution.

Additionally, our model extends beyond the information cascade framework to accommodate diverse epidemic models. Furthermore, their result is limited to expanders, while in our result we obtain error bounds for general network structure.

Turning to the insightful work of Eckles et al. (2022), they focus on optimal seeding under the independent cascade model with a fixed seeding budget. Their strategy is twofold: First, using an oracle that reveals an infection's final spread from a chosen node, they show that small queries achieve an ε-approximation of the optimal seeding solution. Second, where observing edges is costly, they propose a probing algorithm that finds optimal seeds by querying a constant $f$raction of edges. The common thread between our work and theirs is to use small network information for diffusion tasks. However, the specific goal of optimizing seeds is not within the scope of our study.

Another avenue in epidemics on networks focuses on inferring the global network structure from the limited observations of epidemic trajectories. For example, Graham (2008), Goldsmith-Pinkham and Imbens (2013), Netrapalli and Sanghavi (2012) reconstruct networks by fitting parameters with observed data under the assumption that the intrinsic network draws from a stochastic block model (or what they call a linear-in-mean model), an approach later extended to dynamic data by Kim et al. (2014), Drakopoulos and Zheng (2017). In our work, rather than learning network model parameters from existing data, we are interested in data collection to make predictions.

Our data collection algorithm leverages the concept of backward simulation—a method closely related to backward contact tracing, which has been employed in outbreak investigations since at least (Hethcote and Yorke 1984) and revisited during the COVID-19 pandemic (Kojaku et al. 2021, Raymenants et al. 2022). Traditionally, backward tracing is used as a *preventative* tool to identify infection sources and delineate transmission clusters, as demonstrated in the management of diseases such as tuberculosis, HIV, and measles (Lala et al. 2015, Müller and Kretzschmar 2021, Mbivnjo et al. 2022). In our work, we show that data of the same nature can be used for predicting the future course of epidemics. Also, we demonstrate that such retrospective data need not rely solely on realized infections or retrospective interviews. Instead, we propose a theoretical framework wherein the retrospective component is systematically replaced by Monte Carlo simulations of epidemic spread on network structures —without waiting for actual infection events to occur. We provide rigorous theoretical bounds on how this data can be used for prediction, highlighting its potential for epidemic forecasting.

Asymptotic Analysis of Epidemics: Many studies have built rigorous foundation on concentration of time evolution of epidemics under different random graph models, including Erdös-Rényi graphs (Budhiraja et al. 2012, Coppini et al. 2020), configuration models (Janson et al. 2014, Decreusefond et al. 2012), and their dynamic variants (Jacobsen et al. 2018, Ball and Britton 2022, Milewska et al. 2025). Additionally, insights on the duration of epidemics have been illuminated by works like (Bhamidi et al. 2014, Lashari et al. 2021). For a more comprehensive overview of mathematical models of epidemics, the book by (Kiss et al. 2017) serves as a great resource. Many of these random network models meet the tightness and stable neighborhood conditions required by our theorem, thus our concentration results in Theorem 2 are applicable to them.

Most recently, the beautiful works of (Lacker et al. 2023, Ganguly and Ramanan 2024) study a general class of processes on locally convergent graphs, showing that given certain conditions, the epidemics converges to its limit (similar to Theorem 3). While (Lacker et al. 2023) focuses on diffusion processes, (Ganguly and Ramanan 2024) extends the analysis to jump processes, encompassing the SIR model. Their results rely on the stricter assumption of *finitely dissociable graphs* (e.g., those with a bounded maximum degree or unimodular branching processes) to ensure well-defined limits. In contrast, our convergence result (Theorem 3) does not require assumptions on the maximum degree or a specific limiting structure for the graph. Also, our main focus is on the *estimation* of the epidemic using local samples, which is not studied in their result.

2. Model and Main Results

To facilitate a clear presentation, we first present our main results for a classic Susceptible-Infected-Removed (SIR) model. In this model, individuals can be susceptible, infected, or recovered. In the SIR model on networks, individuals are represented as nodes within a graph, denoted as G_n of size

n. The potential transmissions between these individuals are represented as edges. We use $V(G_n)$ and $E(G_n)$ to denote the sets of nodes and edges, respectively. An infected node recovers at a time drawn from an arbitrary recovery time distribution D_R, and while infected, it transmits the disease to its neighbors following a Poisson time process with a fixed rate (denoted by D_I), after which the infected node recovers and becomes immune to reinfection. We assume that each node is initially infected independently with probability $\rho > 0$.

We define the vector $\mathscr{E}_n^{(\rho)}(t) = (S_n^{(\rho)}(t), I_n^{(\rho)}(t), R_n^{(\rho)}(t))$ to represent the state of the epidemic at time t. Each component of this vector—$S_n^{(\rho)}(t)$, $I_n^{(\rho)}(t)$, and $R_n^{(\rho)}(t)$—is a random variable indicating the proportion of susceptible, infectious, and recovered nodes in G_n at time t, respectively. These proportions are calculated as

$$S_n^{(\rho)}(t) = \frac{1}{n} \sum_{v \in V(G_n)} \mathbb{1}\{v \in S \text{ at time } t\}, \qquad I_n^{(\rho)}(t) = \frac{1}{n} \sum_{v \in V(G_n)} \mathbb{1}\{v \in I \text{ at time } t\}, \tag{1}$$

and

$$R_n(t) = \frac{1}{n} \sum_{v \in V(G_n)} \mathbb{1}\{v \in R \text{ at time } t\}. \tag{2}$$

Our results extend beyond the classic SIR model. We prove our results on a general model of epidemics with time-varying infectiousness and heterogeneous initial conditions in Section 2.5.

2.1. Local Estimator for SIR Model

We introduce a local algorithm that uses the information of a small number of individuals in the local neighborhood of the node v to determine its state with respect to the epidemic at any time in the future $t \geq 0$. This is achieved by a backward simulation of the epidemic process.

Given a parameter k and an initial node v as input, the algorithm simulates the backward epidemic process starting from v until it *probes* at most k other nodes in the neighborhood of v. The output is a vector $(S_{k,v}(t), I_{k,v}(t), R_{k,v}(t))_{t \geq 0}$, where $S_{k,v}(t)$, $I_{k,v}(t)$, and $R_{k,v}(t)$ are indicators showing whether node v is susceptible, infectious or recovered at time t under the simulated process.

A more precise illustration of the backward process is as follows: The algorithm requires as input an integer $k > 0$ representing the 'probing' budget, the target node v, the set of initially infected nodes I_0, the recovery time distribution D_R, and the contact time distribution D_I. Our algorithm introduces the notions of 'discovery' and 'probe'. A node is considered 'discovered' when it is first encountered in the simulation. Upon discovery, each new node u is assigned a recovery time r_u sampled from D_R; this recovery time remains fixed for u throughout the simulation.

To 'probe' a node means to explore its connections within the graph G_n. Specifically, when probing a node w, the algorithm queries to obtain w's neighbors, denoted by $N(w)$. For each neighbor $u \in N(w)$, if u has not been discovered before, its recovery time is sampled from D_R (as described above) and a contact time $c_{(u,w)}$ is sampled from D_I. A transmission along (u,w) is considered possible if $c_{(u,w)} < r_u$; in that case, node u is added to the queue for subsequent probing.

The algorithm starts with the target node v in a first-in, first-out probe queue. At each step, it probes (and then removes) the first node in the queue. The simulation progresses as the algorithm continues to probe nodes, incrementally uncovering the structure of G_n and simulating backward infection. This iterative process continues until either the probe queue is empty or the probing budget of k nodes is exhausted.

The output of the algorithm is the estimated time of infection and recovery of the target node v. For this purpose, we use the contact times as edge weight (with weight equal to ∞ where contact time exceeds recovery time) to define a directed distance $\text{dist}_{(T)}(x,y)$ between two nodes x,y as the length of the shortest weighted path from x to y. This distance conceptually represents the time it takes for node x to infect y. Leveraging this construction, the algorithm determines v's infection time by finding the shortest path with respect to $\text{dist}_{(T)}$ from the observed initially infected nodes to v. Here, we assumed the knowledge of each observed node's initial state. Without this knowledge, we can assume the initial state of each node is determined by infecting with probability ρ. Furthermore, the recovery time of node v is obtained by adding r_v to the infection time of v. Finally, the algorithm outputs the state of node v across time. See the details in Algorithm 1 below.

Algorithm 1: Local algorithm – the backward epidemic process

```
Input: Integer $k > 0$, target node $v$, $I_0$, $D_I$, $D_R$.
$Q \leftarrow \{v\}$, $Discovered \leftarrow \emptyset$, $Probed \leftarrow \emptyset$
while $Q$ is not empty and $|Probed| < k$ do
    $w \leftarrow Q.pop()$                          ▷Remove and probe the first node in the queue
    If $w \notin Probed$ for each neighbor $u$ of $w$ do
        $c_{(u,w)} \leftarrow$ sample from $D_I$
        if $u \notin Discovered$ then
            $Discovered \leftarrow Discovered \cup \{u\}$
            $r_u \leftarrow$ sample from $D_R$
        if $c_{(u,w)} < r_u$ then
            $Q.push(u)$
    $Probed \leftarrow Probed \cup \{w\}$
$\mathrm{inf}_v = \mathrm{dist}_{(T)}(Discovered \cap I_0, v)$.   ▷ Shortest transmission path from $Discovered \cap I_0$ to $v$
$\mathrm{rec}_v = \mathrm{inf}_v + r_v$.
$S_{k,v}(t) = \mathbb{1}\{t \leq \mathrm{inf}_v\}$, $I_{k,v}(t) = \mathbb{1}\{\mathrm{inf}_v < t \leq \mathrm{rec}_v\}$, $R_{k,v}(t) = \mathbb{1}\{t > \mathrm{rec}_v\}$.
Output: $(S_{k,v}(t), I_{k,v}(t), R_{k,v}(t))_{t\geq 0}$
```

We choose q independent uniform starting nodes $v_1, \ldots, v_q$, and apply Algorithm 1 to them to estimate the time evolution of the epidemic from these nodes. Assuming that initially each node is infected with probability ρ (so that I_0 is generated accordingly), we define the estimator

$$\hat{S}^{(\rho)}_{q,k,n}(t) = \frac{\sum_{i=1}^{q} S_{k,v_i}(t)}{q} \tag{3}$$

as the fraction of susceptible nodes in the q starting nodes at time t. Similarly, define $\hat{I}^{(\rho)}_{q,k,n}(t)$, $\hat{R}^{(\rho)}_{q,k,n}(t)$, as the fractions of infectious and recovered nodes at time t.

2.2. Efficacy of the Local Estimator

In the following sections, we rigorously examine the error of our local estimators, $\hat{S}^{(\rho)}_{q,k,n}(t)$, $\hat{I}^{(\rho)}_{q,k,n}(t)$, and $\hat{R}^{(\rho)}_{q,k,n}(t)$ in predicting the time evolution of epidemics $S^{(\rho)}_n(t)$, $I^{(\rho)}_n(t)$, and $R^{(\rho)}_n(t)$. We start by presenting an exact bound for the accuracy of the estimator with a predetermined number of queries q and input size k in finite deterministic graphs (see Section 2.2.1). In Section 2.2.2, we extend these insights to random network models, and in Section 2.2.3 we extend our findings to sequences of growing graphs with similar local structures, leveraging the theory of local graph limits.

2.2.1. Finite Deterministic Graphs. Given a fixed deterministic graph G_n, recall that the vector $\mathscr{E}_n^{(\rho)}(t) = (S_n^{(\rho)}(t), I_n^{(\rho)}(t), R_n^{(\rho)}(t))$ represents the epidemic state at time t. Similarly, define $\hat{\mathscr{E}}_{q,k,n}^{(\rho)}(t) = (\hat{S}_{q,k,n}^{(\rho)}(t), \hat{I}_{q,k,n}^{(\rho)}(t), \hat{R}_{q,k,n}^{(\rho)}(t))$ as the estimator vector, obtained from running Algorithm 1 with input k and q independent starting points. We can directly bound the error of the estimator using the following expression:

$$\varepsilon_r(G_n,k) = \frac{1}{n} \sum_{v \in V(G_n)} \mathbb{1}\{|B_r(G_n,v)| > k\},$$

where $B_r(G_n,v)$ is the subgraph of G_n containing all nodes at a graph distance of at most r from v. This expression captures a *tightness* condition on the neighborhood sizes around uniform random nodes.

Theorem 1 **(Local Estimator for a Finite Graph)**. *Let G_n be a deterministic graph of size n. Consider an SIR epidemic in which each node is initially infected with an independent probability of $\rho > 0$. Then for any $t \in [0,\infty]$,*

$$\mathbb{P}\Big(|\mathscr{E}_n^{(\rho)}(t) - \mathbb{E}[\mathscr{E}_n^{(\rho)}(t)]| > \delta\Big) \leq \frac{16}{n\delta^2} + \min_{r,k \geq 1}\Big(\frac{k}{n} + \frac{16\varepsilon_{2r}(G_n,k)}{\delta^2} + \frac{(1-\rho)^r}{\delta}\Big). \tag{4}$$

Further, the error of the estimator is bounded, i.e., for any $t \in [0,\infty]$,

$$\mathbb{P}\Big(|\hat{\mathscr{E}}_{n,q,k}^{(\rho)}(t) - \mathscr{E}_n^{(\rho)}(t)| > \delta\Big) \leq \frac{32(k+1)}{\delta^2 n} + 2e^{-2q\delta^2} + \frac{2}{\delta}\min_{r \geq 0}\big((1-\rho)^r + (1+\frac{16}{\delta})\varepsilon_{2r}(G_n,k)\big). \tag{5}$$

In (4) *and* (5)*, the probability is over both the randomness of the algorithm and the epidemic process.*

Our theorem presents an upper bound on the estimator's error. This bound consists of multiple terms, reflecting the nuanced interplay between various parameters. A main component of our bound depends on the interrelation between the radius r and and the fraction of nodes with r-neighborhood larger than k, expressed as $\varepsilon_r(G_n,k)$. We can bound this term in many scenarios. For example, consider the case that G_n has maximum degree Δ. Then the r neighborhood of the node has at most Δ^r nodes, so we can choose $r < \log_\Delta(k)$ small enough to deduce $\varepsilon_{\lfloor \log_\Delta(k) \rfloor}(G_n,k) = 0$. As a result, the error in (5) can be as small as desired.

What happens if the graph does not have a maximum degree limit, meaning that the maximum degree in G_n may grow quickly with n? Our theorem can still imply that with small k and q the error is small, provided that the size of the local neighborhood of a uniform random node grows slowly. One can ensure such behavior by implementing a constraint on bounding $\varepsilon_r(G_n,k)$ for large enough constant k. We refer to this specific constraint as tightness, elaborated in Definition 1. In Theorem 7, we prove that under the tightness condition, and for any precision δ, there are constant q and k such that the estimator achieves δ-additive error with probability at least $1-\delta$.

Remark 1 (Local estimator of the final size of the epidemic). Since the guarantees of the theorem are independent of t, they also give guarantees for the final size of the infection. ◀

2.2.2. Random Graphs. To determine the error margin of the estimator for random graphs, we begin by illustrating how our algorithm accommodates random network models. To prove the accuracy of the estimator, we require a condition called *Stable Neighborhood Structure* (see Definition 2). This condition ensures that distributions of local network structures are similar across different realizations of the graph. With this condition, we show that for any precision $\epsilon > 0$, there exist constants k_ϵ and q_ϵ with respect to the network size leading to an error of at most ϵ:

Theorem 2 **(Local Estimators for Random Graphs)**. *Let $(G_n)_{n\in\mathbb{N}}$ be a sequence of random graphs satisfying tightness and stable neighborhood structures (Definitions 1 and 2). Then $\mathscr{E}_n^{(\rho)}(t)$ is concentrated,*

$$\left|\mathscr{E}_n^{(\rho)}(t) - \mathbb{E}[\mathscr{E}_n^{(\rho)}(t)]\right| \xrightarrow{\mathbb{P}} 0 \quad as \quad n\to\infty.$$

Furthermore, given any $\epsilon > 0$, there exists constants N,q,k such that for any $n > N$, and any $t\geq 0$,

$$\mathbb{P}\Big(|\hat{\mathscr{E}}_{n,q,k}^{(\rho)}(t) - \mathscr{E}_n^{(\rho)}(t)| > \epsilon\Big) \leq \epsilon. \tag{6}$$

Policy implications of this result arise in scenarios where social planners might lack access to the specifics of the social interactions (even for the few local samples required by Theorem 1). Alternatively, the planner may wish to model the interaction networks and implement a rapid algorithm to predict how an epidemic might unfold. Our result shows that Algorithm 1 can be implemented fast, since it needs to be run on a few nodes. This would enable planners to test and compare different policies efficiently.

Remark 2 (Application to Random Graph Models). The stable local neighborhood and tightness assumptions apply to most sparse random network models. In particular, we show that both conditions hold for graphs converging locally in probability (see Appendix C.3). Thus, our results to configuration models (Dembo and Montanari 2010), sparse inhomogeneous random graphs (including stochastic block models) (Bollobás et al. 2007), preferential attachment models (Berger et al. 2014, Garavaglia et al. 2022), random intersection graphs (Kurauskas 2022, van der Hofstad et al. 2021), random graph models with communities (Trapman 2007, Ball et al. 2010, van der Hofstad et al. 2016), and spatial inhomogeneous random graphs (van der Hofstad et al. 2023), including hyperbolic random graphs (Krioukov et al. 2010, Komjáthy and Lodewijks 2020). For an in-depth overview of network models and their corresponding limits, see van der Hofstad (2024). ◀

2.2.3. Sequence of Growing Graphs. We generalize our previous results using the theory of local graph limits (Benjamini and Schramm 2001, Aldous and Steele 2004). Intuitively, a sequence of (possibly random) graphs $\{G_n\}_{n\in\mathbb{N}}$ is said to have a *local limit in probability* if the empirical distributions of the neighborhoods of randomly sampled nodes converge in probability. The limit is then a probability measure μ on the space $\mathscr{G}_\star$ of rooted, locally finite graphs. We will use (G,o) for a graph G with root o in $\mathscr{G}_\star$. See Section 3.3 for the precise definitions. We prove that epidemics exhibit well-defined limit behavior under local convergence. Further, the final size and the time evolution of an epidemic in finite graphs converge to those on the limit graph.

Theorem 3 **(Convergence of the Epidemic Processes)**. *Let $(G_n)_{n\geq 1}$ be a graph sequence that converges locally in probability to $(G,o)\sim\mu$, where μ is a deterministic probability measure over $\mathscr{G}_\star$. Then, there are functions $\mathscr{E}(t)=(s(t),i(t),r(t))$ such that, for any $t\geq 0$ $\mathscr{E}_n(t)\xrightarrow{\mathbb{P}}\mathscr{E}(t)$, and further, $R_n^{(\rho)}(\infty)/n\xrightarrow{\mathbb{P}} r(\infty)$.*

One implication of this result is that epidemics are essentially a *local property* of the graphs. As a result, it is possible to relate epidemic dynamics across networks with shared local structures but vastly differing scales.

In Theorem 3, the functions $s(t)$, $i(t)$ and $r(t)$ can be expressed in terms of the limit graph:

$$s(t) = \mu(o \in \mathcal{S}^{(\rho)}(t)), \qquad i(t) = \mu(o \in \mathcal{I}^{(\rho)}(t)), \qquad r(t) = \mu(o \in \mathcal{R}^{(\rho)}(t)),$$

where $(\mathcal{S}^{(\rho)}(t), \mathcal{I}^{(\rho)}(t), \mathcal{R}^{(\rho)}(t))$ are the sets of susceptible, infected and recovered nodes for an epidemic on (G,o) started from $\mathcal{R}^{(\rho)}(0) = \varnothing$, and every node is in $\mathcal{I}^{(\rho)}(0)$ independently with probability ρ. Here, the final size of the epidemic can be described by taking the time to infinity, $R_n^{(\rho)}(\infty) = \lim_{t\to\infty} R_n^{(\rho)}(t)$. Equivalently, the final size of the epidemic can also be described as $\mu(o \in \mathcal{R}^{(\rho)}(\infty)) = \mu(\mathscr{C}^{-}(o) \cap \mathcal{I}^{(\rho)}(0) \neq \varnothing)$, where $\mathscr{C}^{-}(o)$ is the set of all nodes reached by the backward epidemic process started from o. As part of the proof of the theorem, we show that the epidemic functions $\mathscr{E}(t) = (s(t), i(t), r(t))$ are well-defined on the limit graph.

Algorithmic insights in the limit: In our proofs, we will show that any sequence of locally convergent graphs satisfies the tightness and stable neighborhood conditions (Definitions 1 and 2). As a consequence, we can put the three theorems together to yield the local estimate of the limit.

THEOREM 4 **(Local Estimation of the Limit)**. *Assume $(G_n)_{n\in\mathbb{N}}$ and the epidemic process satisfy the conditions of Theorem 3. Then given $\epsilon > 0$, there exists constants k q such that for any $t \in [0,\infty]$*

$$\limsup_{n\to\infty} \mathbb{P}\Big(|\hat{\mathscr{E}}^{(\rho)}_{q,k,n}(t) - \mathscr{E}(t)| > \epsilon\Big) \leq \epsilon. \tag{7}$$

2.3. Robustness to Weighted Node Selection

So far, we focused on the setting where the input nodes $v_1, \ldots v_q$ for algorithm is sampled from a uniform distribution. In some applications, however, it may be necessary or preferable to select nodes with unequal sampling weights. For example, in practice, data might disproportionately cover a specific demographic or community. One way to capture such non-uniform sampling is to assume that each node i is chosen with probability $p_i = c_i/n$, where $(c_i)_{i\in[n]}$ is a set of positive weights summing to n, where $[n] = \{1, 2, \ldots, n\}$. Intuitively, c_i represents how likely it is that a policymaker will choose node i for data collection. In this section, we show how to adapt our local estimator to account for these $(p_i)_{i\in[n]}$, and we prove that the same main structural results still hold: the estimator has a small (and explicitly bounded) bias, and its variance remains controlled by the tightness of the graph.

Horvitz–Thompson Local Estimator. Let $\{v_1,\dots,v_q\}$ be q independent draws of nodes from the weighted distribution (p_i). Run the backward simulation (Algorithm 1) as before, obtaining indicators $S_{k,v_j}(t)$, $I_{k,v_j}(t)$, $R_{k,v_j}(t)$. To build the estimator, we need to incorporate the weight $1/p_{v_j}$. Thus, define the estimator

$$\widehat{S}_{q,k,n,\vec{p}}(t) \;=\; \frac{1}{q}\sum_{j=1}^{q}\frac{S_{k,v_j}(t)}{n\,p_{v_j}},$$

and similarly $\widehat{I}_{q,k,n,\vec{p}}(t)$, $\widehat{R}_{q,k,n,\vec{p}}(t)$. This is also known as Horvitz–Thompson estimator Horvitz and Thompson (1952).

We will show that the bias of $\widehat{S}_{q,k,n,\vec{p}}(t)$ matches that of the uniform local estimators (i.e., where $p_i = 1/n$). In particular, it is bounded by $(1-\rho)^r$, provided k is large enough to cover the r-balls of typical nodes (cf. Lemma 1). The variance is governed a similar tightness-based bound (Lemma 2), with an additional factor accounting for $\frac{\max_{i\in[n]} p_i}{\min_{i\in[n]} p_i}$. This is formalized in the following result:

Theorem 5 **(Weighted Initial Sampling)**. *Let $\widehat{S}_{q,k,n,\vec{p}}(t)$ be the weighted estimator described above, constructed by sampling q nodes i.i.d. with probabilities $p_i = c_i/n$. Then, for any given graph G_n of size n,*

$$\sup_{t\geq 0}|\mathbb{E}\Big[\widehat{S}_{q,k,n,\vec{p}}(t)\Big]-\mathbb{E}\Big[S_n(t)\Big]|\leq \min_r\big((1-\rho)^r+\varepsilon_r(G_n,k)\big),\qquad \textit{and}$$

$$Var\big(\widehat{S}_{q,k,n,\vec{p}}(t)\big)\leq \frac{1}{\min_i np_i}\Big(\frac{1}{q}+\frac{1}{n}\Big)+\Big(\frac{\max_{i\in[n]} p_i}{\min_{i\in[n]} p_i}\Big)^2\varepsilon_{2k}(G_n),$$

where $\varepsilon_{2k}(G_n)=\frac{1}{n^2}\left|\{(u,v)\in V(G_n)\times V(G_n)\colon \mathrm{dist}_{G_n}(u,v)\leq 2k\}\right|$. Moreover, if the graphs $(G_n)_{n\in\mathbb{N}}$ also satisfy the condition of Theorem 2, then for any given $\epsilon>0$, there exists k_ϵ such that for any q and large enough n,

$$\sup_{t\geq 0}|\mathbb{E}\Big[\widehat{S}_{q,k_\epsilon,n,\vec{p}}(t)\Big]-\mathbb{E}\Big[S_n(t)\Big]|\leq\epsilon,\qquad \textit{and}\qquad Var\big(\widehat{S}_{q,k_\epsilon,n,\vec{p}}(t)\big)\leq\epsilon\Big(\frac{\max_{i\in[n]} p_i}{\min_{i\in[n]} p_i}\Big)^2+\frac{1}{q\min_i np_i}.$$

Note that the variance is essentially controlled by the fraction of node pairs whose distance is at most $2k$; as we show in the proof of Theorem 1, this term is upper bounded by $\varepsilon_r(G_n, k)$. This result (with its proof in Appendix E) shows that there exists a natural tradeoff: a larger k decreases bias by better approximating the global epidemic dynamics, but it may simultaneously increase the variance, due to the higher chance of overlapping neighborhoods of two samples[1]. The optimal choice of k thus depends on the tightness properties of the graph. Note that also this result applies to the uniform sampling case (i.e., when $c_i = 1$ so that $p_i = 1/n$).

2.4. Robustness to Misspecified Distributions

In practical scenarios, the true infection and recovery distributions may not be perfectly known. Instead, one often relies on estimates or approximations of these distributions derived from empirical data. Let $\widetilde{D}_I$ and $\widetilde{D}_R$ be such approximate distributions used in place of the true D_I and D_R. We assume that they lie within bounded total variation (TV) distance of the real distributions, i.e.,

$$d_{\mathrm{TV}}(D_I, \widetilde{D}_I) \le \epsilon_I \quad \text{and} \quad d_{\mathrm{TV}}(D_R, \widetilde{D}_R) \le \epsilon_R,$$

where $d_{\mathrm{TV}}(\mu, \nu) := \sup_A \big|\mu(A) - \nu(A)\big|$ denotes the total variation distance between two probability measures μ and ν, and ϵ_I and ϵ_R are small constants.

Our goal is to show that the epidemic process (as well as its estimator) remains stable under these bounded perturbations. Our results demonstrate that the error introduced by using misspecified distributions $\widetilde{D}_I$ and $\widetilde{D}_R$, instead of the true distributions D_I and D_R, can be bounded by the total variation distances $\epsilon_I = d_{\mathrm{TV}}(D_I, \widetilde{D}_I)$ and $\epsilon_R = d_{\mathrm{TV}}(D_R, \widetilde{D}_R)$. The next result with its proof in Appendix F formalizes this.

THEOREM 6. *Let G_n be a deterministic graph of size n. Also, let $\tilde{\mathscr{E}}^{(\rho)}_{n,q,k}(t)$ the estimator vector, obtained from running Algorithm 1 with $\widetilde{D}_I$ and $\widetilde{D}_R$ used as contact times. Then*

$$d_{TV}\Big(\tilde{\mathscr{E}}^{(\rho)}_{n,q,k}(t), \hat{\mathscr{E}}^{(\rho)}_{n,q,k}(t)\Big) \le k^2(\epsilon_I + \epsilon_R)(1 - \varepsilon_r(G_n, k)) + \varepsilon_r(G_n, k).$$

[1] In the proof of Theorem 1, we provide bounds on $\varepsilon_r(G_n)$ based on $\varepsilon_r(G_n, k)$.

Note that the error bound in Theorem 6 is uniform in time, meaning that the robustness guarantee holds for all $t \geq 0$. Moreover, the term $\varepsilon_r(G_n, k)$ captures the error due to truncating the local neighborhood and is controlled by our tightness condition. Overall, this result demonstrates that our method is robust to moderate misspecifications in epidemic parameters, making it well suited for practical applications where such parameters must be estimated from data.

2.5. Generalization to Other Models

The core results presented in this work, specifically the convergence of epidemic processes in Theorem 3, generalize in several important ways. In this section, we explore the applicability of our results to various epidemic models and starting configurations, providing a more comprehensive picture of the theorem's reach.

2.5.1. General Epidemics We introduce a generalized epidemic model with time-varying infectiousness that would apply to SI, SEIR, or different variations of it, as well as, to intervention strategies such as vaccination.

Time-varying infectiousness: We explore a model where a node's infectiousness varies over time, accommodating stages like exposure or fluctuating infectiousness levels (see Figure 1). This model distinguishes two timescales: 1) the *epidemic timescale* tracking disease progression network-wide, and 2) *node-specific timescales* beginning when a node becomes infected.

In this model, nodes are either susceptible or occupy a disease state from $\mathscr{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$. This set describes the m sequential states a node undergoes after infection, starting with $\mathcal{D}_1$ and subsequently moving to $\mathcal{D}_2$, and so forth. Once a node in the disease state transmits the disease to its susceptible neighbor, that neighbor enters the $\mathcal{D}_1$ state, starting its node-specific timescale.

This timescale $[0, \infty]$ is partitioned into m intervals $[t_0, t_1), [t_1, t_2), \cdots [t_{m-1}, \infty]$, where the interval $[t_{i-1}, t_i)$ corresponds to the state $\mathcal{D}_i$. The transition times $t_1 \leq t_2 \leq \ldots \leq t_m$, are drawn from a distribution $\tau_v : \{t_1, t_2, \ldots, t_m\} \to \mathbb{R}_+^m$. For example, Figure 1 shows states labeled as Exposed, Infectious, Quarantine, and Recovered. The figure also shows how a node's infectiousness varies over time, which we describe next.

The infectiousness of node v is determined by a probability density function $\beta_v : [0,\infty] \to \mathbb{R}_+$[2]. We sample (τ_v, β_v) from a joint probability distribution P_β, which couples the duration of each disease state with its corresponding infectiousness. Further, we assume that P_β depends on the local network structure, i.e., there exists an integer $\ell > 0$ such that β_v and τ_v are drawn from $P_\beta(B_\ell(G,v))$.

Then, the epidemic progresses as follows. First, for each node v, draw β_v and τ_v from $P_\beta(B_\ell(G,v))$. Then, for each neighbor of v, draw its transmission times independently from β_v. Initially, each node is in state $\mathcal{D}_1$ with probability $\rho > 0$. Incorporating Algorithm 1 into this epidemic model is similar: for each node v, sample β_v and τ_v from P_β, and then draw transmission times from β_v, yielding a weighted directed graph, from which the backward edges can be identified.

Our main results apply to this general epidemic model. As before, we define $\mathscr{E}_n(t)$ as the fraction of nodes in each state of the epidemic (susceptible and disease states $\mathscr{D}$) at time t in a finite graph G_n, $\mathscr{E}(t)$ for the limit graph, and $\hat{\mathscr{E}}_{q,k,n}(t)$ as the estimator using q queries and k as input.

Corollary 1 **(Convergence of Epidemics with Time-varying Infectiousness)**. *Let $(G_n)_{n\geq 1}$ satisfy the conditions of Theorem 3. Consider an epidemic model with time-varying infectiousness as above. Then the epidemic concentrates for any $t \geq 0$, $\mathscr{E}_n(t) \xrightarrow{\mathbb{P}} \mathscr{E}(t)$. Further for any given $\delta > 0$, there exists constants k, q such that, for all $t \in [0,\infty]$*

$$\limsup_{n\to\infty} \mathbb{P}\Big(|\hat{\mathscr{E}}_{q,k,n}(t) - \mathscr{E}(t)| > \delta\Big) \leq \delta. \tag{8}$$

The time-varying infection model is an adaptable framework that accommodates various epidemic models, including additional states like an exposed period and vaccination.

Example 1 (Epidemics with Vaccination). Consider a scenario where specific nodes within a network receive vaccinations based on a locally defined probability function, guaranteeing their immunity against the disease. This scenario can be represented using the time-varying infectiousness epidemic model, where $\mathscr{D} = \{I, R, V\}$ has three states of Infectious, Recovered, and Vaccinated.

[2] The precise condition is that $\beta_v(t) \geq 0$ for all t $\in [0,\infty)$, and that β_v integrates to one, i.e., $\int_0^\infty \beta_v(t)\,\mathrm{d}t = 1$.

In this model, for vaccinated nodes, P_β allocates a β_v with zero density probability and τ_v that defines the time of vaccination at $t_3 = 0$ (as a result, $t_1 = t_2 = 0$). For all other nodes, P_β assigns an infectiousness density function as usual. ◀

Remark 3 (Generalizations for finite graphs). While this section primarily focuses on graph sequences with a local limit, the findings extend to finite graphs, under analogous tightness and stable neighborhoods discussed before in Section 2.2.1. The only difference is that we need to add a notion of 'marks' (see Section 3.4). We have chosen not to include these cumbersome notational adjustments to keep the primary exposition clear. ◀

2.5.2. General Starting Configurations. In our main results so far, we assumed that each node is initially infected independently with a probability $\rho > 0$. We can generalize our results to heterogeneous initial states, where a node's initial state is drawn from a probability distribution depending on its local neighborhood, for example, the node degree. As another example, the initial state could be determined by the PageRank, which can be approximated by local network structures (Garavaglia et al. 2020). To formalize this, we assume there exists a function P_ℓ, where given a node's ℓ-neighborhood $B_\ell(G,o)$ as input, $P_\ell(B_\ell(G,o))$ provides a probability distribution on the initial states of the node o, whether S, or I.

We further need a second condition ensuring the presence of an initially infected individual within any sufficiently large path starting from a uniformly random node. We refer to this as the *locally reachable property.* To formalize this, for any vertex $v \in V(G_n)$ and any integer $r \geq 1$, let $\mathrm{Path}_r(v) := \{\text{all paths of length } r \text{ starting at } v\}$, where a *path of length r* is defined as any sequence $(v_0, v_1, \ldots, v_r)$ with $v_0 = v$, $v_0, v_1, \ldots, v_r$ all distinct, and there exists an edge between v_i and v_{i+1} for $0 \leq i \leq r-1$. Then the sequence $(G_n)_{n\geq 0}$ with the initial states drawn from P_ℓ is said to be *locally reachable* if, for any $\delta > 0$,

$$\lim_{r\to\infty} \limsup_{n\to\infty} \mathbb{P}\Big(\sup_{\gamma \in \mathrm{Path}_r(v)} \mathbb{P}_{P_\ell}\big(\gamma \cap I_0 = \emptyset \mid G_n\big) \geq \delta \Big) = 0, \tag{9}$$

where the outer probability is taken over the randomness of the graph G_n and the uniform choice of the starting node v, and the inner probability is with respect to the randomness of the initial infection configuration I_0. This property obviously holds for the case of independent initial infection with probability ρ, since the chance that the r-neighborhood of any node does not encounter I_0 is $(1-\rho)^r$, which goes to 0 with $r \to \infty$. More generally, we have the following result:

COROLLARY 2 **(General Starting Configuration)**. *Let $(G_n)_{n\geq 1}$ satisfy the conditions of Theorem 3. Consider a SIR epidemic, where the starting infections are locally reachable* (9)*, and that there exists some ℓ such that the initial conditions are specified by a strictly local function P_ℓ based on ℓ-neighborhoods as defined above. Then the conclusions of Theorems 3 and 4 hold.*

Here, we consider the initial states when we define the epidemic in the limit $\mathscr{E}(t) = (s(t), i(t), r(t))$, i.e, we equipped the measure μ over rooted graphs $\mathscr{G}_\star$ with first drawing the rooted graph (G, o) and then the initial conditions P_ℓ, and denoted it as μ_Ξ. See details in Section 3.4.

EXAMPLE 2 (APPLICATION TO VIRAL MARKETING). Companies often use degree-based seeding, targeting highly connected individuals, to maximize the information spread about a new product. This raises the question of how information spreads under a given seeding strategy - a topic studied under different random graph models (see, e.g., Manshadi et al. (2020), Akbarpour et al. (2018)). Recognizing that degree-based targeting satisfies the local reachable property, a marketing platform can apply our local estimator to predict information spread by observing only a few nodes. ◀

3. Conditions on Local Graph Structures

This section outlines a hierarchy of local graph conditions, each extending the previous one.

3.1. Graphs with Tight Neighborhood Sizes

The first condition is the notion of *tightness*, which requires that in a sequence of graphs, the number of nodes within a fixed radius of any uniformly chosen node is bounded. Recall that $B_r(G, o)$ is the subgraph of (G, o) containing all nodes within graph distance r of o.

DEFINITION 1 (GRAPHS WITH TIGHT NEIGHBORHOOD SIZES.). Let $(G_n)_{n\in\mathbb{N}}$ be a sequence of graphs with $|V(G_n)| = n$, and let $\varepsilon_r(G_n, k)$ be the empirical probability that $B_r(G_n, v)$ contains more than k nodes when v is chosen uniformly at random, i.e.,

$$\varepsilon_r(G_n, k) = \frac{1}{n} \sum_{v\in V(G_n)} \mathbb{1}\{|B_r(G_n, v)| > k\}. \tag{10}$$

We say that the sequence of graphs has *tight neighborhood sizes* if for all $r<\infty$ and all $\delta>0$ there exists $k<\infty$ such that for all n large enough $\varepsilon_r(G_n,k)\leq\delta$. If the graph G_n is itself random, then we say it has tight neighborhood sizes if $\mathbb{P}(\varepsilon_r(G_n,k)\leq\delta)\geq 1-\delta$. ◀

3.2. Graphs with Stable Neighborhood Structures

We extend our constraints to random graphs by introducing the *Stable Neighborhood Structure* condition, which ensures that the empirical distribution of local network structures is similar in different realizations of the random network.

To define this rigorously, we need to define the concept of a 'rooted graph,' which we will use in the following sections as well. A *rooted graph* is a pair (G,o) where $G=(V(G),E(G))$ is a graph with nodes in $V(G)$ and edges in $E(G)$, and $o\in V(G)$ is a specific node. The graphs (G_1,o_1) and (G_2,o_2) are *isomorphic*, denoted as $(G_1,o_1)\simeq(G_2,o_2)$, if there exists a bijection $\phi\colon V(G_1)\mapsto V(G_2)$ such that $\phi(o_1)=o_2$ and $u,v\in E(G_1)$ if and only if $\phi(u),\phi(v)\in E(G_2)$. Also, define $P_r^{(G_n)}(H^\star)=\frac{1}{|V(G_n)|}\sum_{v\in V(G_n)}\mathbb{1}\{B_r(G_n,v)\simeq H^\star\}$ as the probability that the r-ball neighborhood of a uniform random node in G_n is isomorphic to $H^\star$. For random graphs of size n, let $p_r^{(n)}(H^\star)=\mathbb{E}\big[P_r^{(G_n)}(H^\star)\big]$ represent the mean of this probability across all random realizations of G_n on graphs of size n.

Definition 2 (Stable Neighborhood Structure). Let $(G_n)_{n\in\mathbb{N}}$ be a sequence of (possibly random) graphs with $|V(G_n)|=n$. We say that the sequence of graphs has a *Stable Neighborhood Structure* if for all $r<\infty$, all $\delta>0$, and any rooted graph $H^\star$, as $n\to\infty$,

$$\mathbb{P}\Big(|P_r^{(G_n)}(H^\star)-p_r^{(n)}(H^\star)|\geq\delta\Big)\xrightarrow{\mathbb{P}}0.$$ ◀

Stable neighborhood structure ensures that different random samples for similar-sized networks lead to the same empirical distribution of local neighborhood distribution, even if these statistics change for different values of n. To analyze the asymptotics of graphs, a stronger requirement is needed. Not only must we ensure consistent empirical distributions for networks of similar size, but these distributions must also be asymptotically independent of the size of the graph n, leading to the next concept: local convergence in probability.

3.3. Local Convergence in Probability

The foundational framework of local weak convergence was initiated independently by Aldous and Steele (2004) as well as by Benjamini and Schramm (2001). For a more comprehensive treatment, readers are directed to Bordenave (2016) or (van der Hofstad 2024, Chapter 2).

At a high level, a sequence of graphs $(G_n)_{n\in\mathbb{N}}$ is said to exhibit local convergence if the empirical distribution governing the neighborhoods of randomly chosen nodes approximates a certain limit distribution. To define this rigorously, we must introduce a metric on the space of rooted graphs.

We denote the space of (potentially infinite) connected rooted graphs as $\mathscr{G}_\star$, where two rooted graphs are considered equivalent if they are isomorphic. Therefore, $\mathscr{G}_\star$ consists of equivalence classes of rooted graphs modulo isomorphism. This space of rooted graphs, $\mathscr{G}_\star$, can be endowed with a metric structure denoted as $d_{\rm loc}$. The metric $d_{\rm loc}$ between two rooted graphs (G_1,o_1) and (G_2,o_2) is

$$d_{\rm loc}((G_1,o_1),(G_2,o_2)) = \frac{1}{1+\inf_k\{k : B_k(G_1,o_1) \not\simeq B_k(G_2,o_2)\}}.$$

Note that this metric endows $\mathscr{G}_\star$ with the natural σ-algebra of Borel sets, allowing us in particular to consider measures μ on $\mathscr{G}_\star$.

Definition 3 (Local convergence in probability). Let μ be a measure on $\mathscr{G}_\star$. We define the concept of local convergence in probability for a sequence of graphs $(G_n)_{n\geq 1}$ to a limit $(G,o)\sim\mu$ as follows: For every $r\geq 0$ and $H^\star\in\mathscr{G}_\star$,

$$\frac{1}{|V(G_n)|}\sum_{v\in V(G_n)} \mathbb{1}\{B_r(G_n,v)\simeq H^\star\} \xrightarrow{\mathbb{P}} \mu(B_r(G,o)\simeq H^\star). \tag{11}$$

This definition implies that the proportions of subgraphs in the random graph G_n converge in probability towards those prescribed by μ. Other notions of local convergence, such as local weak convergence, where the focus shifts to the convergence of expectations, and local almost sure convergence, which considers almost sure convergence, are related. However, for our current purposes, local convergence in probability is the most convenient choice, particularly due to its implication that the neighborhoods of two uniformly chosen nodes do not overlap; see (van der Hofstad 2024, Corollary 2.18) for further details.

3.4. Mark Local Convergence

In our framework, the notion of *marks* will play a pivotal role. These marks correspond to attributes associated with the infection and recovery times of the epidemic, as well as the initial states of the nodes. We assume the marks are defined on some complete separable metric space Ξ, accompanied with the metric d_Ξ. We define graph marks $\mathcal{M}(G) = ((M(v))_{v\in V(G)}, (M(v,u))_{(u,v)\in E(G)})$ to annotate G with the marks associated with both nodes and edges, where $M(v), M(v,u) \in \Xi$. Similarly, a marked rooted graph $(G, o, \mathcal{M})$ is a rooted graph (G, o) with the corresponding marks. Here, edges are considered as directed with (u,v) showing the direction from u to v. This distinction is particularly relevant in our exploration of epidemics, where the traversal time along a directed edge (v,u) may differ from that of (u,v).

For two rooted marked graphs $(G_1, o_1, \mathcal{M}_1)$ and $(G_2, o_2, \mathcal{M}_2)$, we define ϵ-marked isomorphism as a standard graph isomorphism (ignoring marks) together with the requirement that the distances between the corresponding marks are bounded by ϵ. We denote this relation by $\overset{\epsilon}{\simeq}$, defined as,

$$\begin{aligned}(G_1, o_1, \mathcal{M}_1) \overset{\epsilon}{\simeq} (G_2, o_2, \mathcal{M}_2) := \mathbb{1}\Big(& d_{\mathscr{G}_\star}\big((G_1, o_1), (G_2, o_2)\big) < \epsilon, \text{ and } \exists \pi \text{ such that} \\ & \max_{u \in V(G_1)} d_\Xi\big(\mathcal{M}_1(u), \mathcal{M}_2(\pi(u))\big) < \epsilon \\ & \max_{(u,v)\in E(G_1)} d_\Xi\big(\mathcal{M}_1(u,v), \mathcal{M}_2(\pi(u,v))\big) < \epsilon\Big)\end{aligned}$$

where π runs over all isomorphisms between the $1/\epsilon$-rooted neighborhoods of (G_1, o_1) and (G_2, o_2). Given this notation, local convergence can be generalized to the case of marked graphs.

Definition 4 (Marked Local Convergence in Probability). The sequence $(G_n, \mathcal{M}_n)_{n\geq 1}$ converges locally in probability to $(G, o, \mathcal{M}) \sim \mu_\Xi$, when for any $\epsilon > 0$, and any marked rooted graph $(H^\star, \mathcal{M}^\star)$,

$$\frac{1}{|V(G_n)|} \sum_{o_n \in V(G_n)} \mathbb{1}\{(G_n, o_n, \mathcal{M}_n) \overset{\epsilon}{\simeq} (H^\star, \mathcal{M}^\star)\} \xrightarrow{\mathbb{P}} \mu_\Xi\big((G, o, \mathcal{M}) \overset{\epsilon}{\simeq} (H^\star, \mathcal{M}^\star)\big). \tag{12}$$

This definition implies that a sequence of marked graphs converges if the empirical frequencies of subgraphs, allowing an ϵ tolerance for errors in the marks, converges to a limiting distribution.

Given a finite graph G, we represent the probability distribution on $\mathcal{M}(G)$ as $P_\Xi(\cdot|G)$. If (G,o) is a locally finite graph with a root node o, the notation $P_\Xi(\cdot|(G,o))$ is employed. In this paper, μ_Ξ is constructed by initially sampling $(G,o)\sim\mu$ and subsequently assigning marks using the measure $P_\Xi(\cdot|(G,o))$. Next, we will detail how Ξ and P_Ξ are defined in the context of SIR, SIR with general starting configuration, and epidemics with time-varying infectiousness.

In the context of SIR epidemics, the mark space is denoted as $\Xi=[0,\infty]\times\{S,I,R\}$. The first component of Ξ corresponds to the transmission time for edges and the recovery time for nodes. The second component, which is relevant only to the nodes, represents the initial state of the node. We let the metric $d_\Xi((t,X),(t',X'))=|t-t'|+\mathbb{1}\{X\neq X'\}$ be the discrete metric for the second component and the metric L_1 for the first one. When defining the probability distribution $P_\Xi(\cdot|(G,o))$, the first component samples infection times for each edge from the distribution D_I, and recovery time for each node from D_R. For the second component of marks (which is independent of the first component), each node has an initial state of I with probability ρ, and if not, it is marked as S.

For epidemics with a general starting point, the space of marks remains unchanged as $\Xi=[0,\infty]\times\{S,I,R\}$. Also, the probability distribution P_Ξ on the first component corresponds to the time of transmission, and recovery stays as before. The distinction arises in the second component, representing the epidemic's initial state. In this scenario, we use a function P_ℓ that maps rooted graphs with a radius of at most ℓ to probability distributions over the states $\{S,I,R\}$. With this assumption, the mark of a node depends only of its ℓ-neighborhood, i.e., $P_\Xi(M(o)|(G,o))=P_\Xi(M(o)|B_\ell(G,o))$.

In the context of epidemics with time-varying infectiousness, we refine the definition of marks associated with nodes. Specifically, the node marks $(M(v))_{v\in V(G)}$ now encompass not only the initial state of node – indicating whether it is in a disease state or susceptible– but also the density functions β_v and τ_v. Let $\mathscr{B}$ be the space of pairs of density functions such as (β_v,τ_v), where $\beta_v:[0,\infty]\to\mathbb{R}_+$ is a probability density of transmission time from v to its neighbors and

$\tau_v : [t_1, \ldots, t_m] \to \mathbb{R}_+^m$ shows transition times between different disease states of the node. Then the marks on nodes take values in $\mathscr{B} \times \{0,1\}$, with the first component identifying (β_v, τ_v) drawn from P_β, and the second component indicating whether the initial state of the node is susceptible or a disease state. The metric space for $\mathscr{B}$ is the L^1 norm on functions, and, for the second component, it is the L^1 norm, which makes $\mathscr{B}$ separable and complete. In addition, edge marks, represented as $M(v,u)$, are drawn independently from β_v, indicating the transmission time from v to u. So, the space of marks for edges is $\mathbb{R}_+$. As a result, the space of marks on the graph is the union of node marks and edge marks $\Xi = \mathscr{B} \times \{0,1\} \cup \mathbb{R}_+$. Further, P_Ξ is defined by first drawing node marks and then edge marks as described above.

4. Proof Outline

In this section, we outline the main ideas of the proofs. We use a second-moment argument showing that truncating the epidemic at a constant radius approximates the epidemic on the entire graph.

4.1. Proofs for Finite Deterministic Graphs

We start by proving that running the epidemic within the r-ball of each node, rather than across the entire graph, results in an outcome that concentrates around the expected time evolution of epidemics. This idea is equivalent to running Algorithm 1 for n queries with each node as a starting point, then selecting a sufficiently large k to cover the r-ball of each node. Using a second-moment argument, we will bound both the mean and variance of the truncation.

To formalize our approach, we introduce the notation $T^{(r)}$. This function maps a rooted marked graph $(G, o, \mathcal{M}(G))$ to the infection time of o under the assumption that the network is confined to $B_r(G,o)$. Formally, $T^{(r)}$ assigns a non-negative number to a rooted marked graph $(G, o, \mathcal{M}(G))$, which is equal to the length of the shortest path from the set of initially infected nodes in $B_r(G,o)$ to the root o (with respect to the weighted distance $\text{dist}_{(T)}(\cdot,\cdot)$ defined in Section 2). When the graph is given in the context, we use $T^{(r)}(o)$ as a short form of $T^{(r)}(G, o, \mathcal{M}(G))$.

Our goal is to approximate $S_n^{(\rho)}(t)$ by a sum of functions defined on balls of radius r, namely

$$S_{n,r}^{(\rho)}(t) = \frac{1}{n} \sum_{v \in V(G_n)} \mathbb{1}\{t < T^{(r)}(v)\}. \tag{13}$$

Similarly one can define $I^{(\rho)}_{n,r}(t)$, $R^{(\rho)}_{n,r}(t)$, and the vector $\mathscr{E}^{(\rho)}_{n,r}(t) = (S^{(\rho)}_{n,r}(t), I^{(\rho)}_{n,r}(t), R^{(\rho)}_{n,r}(t))$. The following two lemmas bound the first and second moment of this truncation:

Lemma 1 **(First Moment)**. *For a given finite graph G_n,*

$$\mathbb{E}_{\Xi}\left[\sup_{t\geq 0}\left|\mathscr{E}^{(\rho)}_{n}(t) - \mathscr{E}^{(\rho)}_{n,r}(t)\right|\right] \leq (1-\rho)^r,$$

where the expectation is with respect to the epidemic and the random set of initially infected nodes.

The proof is based on showing that the shortest path, in terms of $\mathrm{dist}_{(T)}$, from a node to the set of initially infected nodes traverses only a limited number of graph nodes. This is because the initially infected nodes are 'locally reachable', and the likelihood of not encountering them reduces geometrically (see Appendix A.1).

Subsequently, we bound the variance of this approximation for the second moment. The main idea is that the local epidemic processes (which define $T^{(r)}$) for nodes separated by more than distance $2r$ are independent. Consequently, we can bound the variance by counting node pairs separated by over-distance r. This quantity is denoted as

$$\varepsilon_r(G_n) = \frac{1}{n^2}\left|\{(x,y) \in V(G_n) \times V(G_n) \colon \mathrm{dist}_{G_n}(x,y) \leq r\}\right|,$$

where $\mathrm{dist}_{G_n}(x,y)$ is the graph distance of x and y in G_n. The detail of the proof appears in Appendix A.2. Later, we provide bounds on $\varepsilon_r(G_n)$ based on $\varepsilon_r(G_n,k)$ (defined for Theorem 1).

Lemma 2 **(Second Moment)**. *Let G_n be a deterministic graph with n nodes. Then*

$$\sup_{t\geq 0}\mathrm{Var}_{\Xi}(S^{(\rho)}_{n,r}(t)) \leq \frac{1}{n} + \varepsilon_{2r}(G_n),$$

where the variance is over the randomness of the epidemic process.

Note that the function $S^{(\rho)}_{n,r}$ is equal to the estimator $\hat{S}^{(\rho)}_{n,r,n}$ when making n queries with an input of r to Algorithm 1. Further, Lemmas 1 and 2 provide the foundation for establishing the concentrations of $S^{(\rho)}_{n,r}(t)$ and $S^{(\rho)}_{n}(t)$. To conclude the proof of Theorem 1, the main step is to determine the accuracy of q queries $\hat{S}^{(\rho)}_{q,r,n}$ to approximate $S^{(\rho)}_{n,r}$. For this purpose, first, we condition on the epidemic process to bound the error of choosing q using standard concentration arguments. Then, to get the overall accuracy of the estimator, we use the variance bound in Lemma 2 to control the estimator's value across different realizations of the epidemic process. See Appendix A.3.

Theorem 1 provides explicit bounds on the estimator's error for a given graph. This bound can be made as narrow as desired for graphs that meet the 'tightness' condition in Definition 1. The following theorem formalizes this – even with a constant number of queries – the estimator closely match the true time evolution of epidemics, $(S_n^{(\rho)}(t), I_n^{(\rho)}(t), R_n^{(\rho)}(t))$, to any chosen precision:

Theorem 7 **(Local Estimation of Tight Graphs)**. *Let $(G_n)_{n\in\mathbb{N}}$ be a sequence of graphs with tight neighborhood sizes. Then $\big|\mathscr{E}_n^{(\rho)}(t) - \mathbb{E}[\mathscr{E}_n^{(\rho)}(t)]\big| \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$. Furthermore, given any $\delta > 0$, there exists constants N, q, k such that for any $n > N$ and $t \geq 0$, $\mathbb{P}\Big(|\hat{\mathscr{E}}_{q,k,n}^{(\rho)}(t) - \mathscr{E}_n^{(\rho)}(t)| > \delta\Big) \leq \delta$.*

In Appendix A.5, we present an example emphasizing the importance of the tightness condition, illustrating that in structures like the star graph, even the final infection size does not concentrate.

4.2. Proofs for Finite Random Graphs

The proof largely mirrors that of deterministic graphs, employing a similar second-moment argument. For the first moment, Lemma 1 suffices as we can use linearity of expectation to get convergence of first-moment with the randomness of G_n taken into account. However, for bounding the variance as in Lemma 2, we must extend our result to accommodate the randomness of the neighborhood structure. Definition 2 ensures a small variance between the expected time evolution of epidemics across different network realizations as it is formalized in Lemma 3. See Appendix A.2. Then, the proof of the Theorem 2 follows the exact steps of Theorem 1, which appears in Appendix B.

Lemma 3 **(Local Estimation of Random Graphs - Second Moment)**. *Let $(G_n)_{n\geq 1}$ be a sequence of (possibly random) graphs with tight and stable neighborhood structures (see Definitions 1 and 2). Then for any given $\delta > 0$, and large enough n, $\sup_{t\geq 0} \mathrm{Var}(S_{n,r}^{(\rho)}(t)) \leq \delta$, where the variance is over the randomness of the epidemic process and the graph G_n.*

4.3. Proofs for Growing Graphs

We establish the local convergence of the epidemic in three steps. First, as in Theorem 2, we prove that an epidemic restricted to a constant radius ball around nodes (represented as $\mathscr{E}_{n,r}(t)$ in (13)) concentrates around the epidemic spanning the entire graph $\mathscr{E}_n(t)$. The second step is the local approximation of the limit graph with a similar truncation. This approach mirrors the method used for the finite graph, which we detail in Lemma 4. The final stage (Lemma 5) ensures the convergence of the truncated epidemic in the finite graph to that of the limit graph.

Starting with the local approximation of the epidemic in the limit, recall the definition of $T^{(k)}(o, G, \mathcal{M}(G))$ from Section 4.1. For the limit graph, define $s_k(t) = \mu_\Xi\Big(\mathbb{1}\{t < T^{(k)}(G, o, \mathcal{M}(G))\}\Big)$. Similarly, $i_k(t)$, and $r_k(t)$ can be defined. Then we can extend Lemma 1 for the limit graph.

Lemma 4 **(Local Approximation of the Limit)**. *For any (deterministic) measure μ on $\mathscr{G}\star$, and any integers k and k',*

$$\mu_\Xi\left[\sup_{t\geq 0}\left|\mathbb{1}\{t < T^{(k)}(G, o, \mathcal{M}(G))\} - \mathbb{1}\{t < T^{(k')}(G, o, \mathcal{M}(G))\}\right|\right] \leq (1-\rho)^{\min\{k,k'\}}.$$

Thus, $s(t) = \lim_{k\to\infty} s_k(t)$, $i(t) = \lim_{k\to\infty} i_k(t)$ and $r(t) = \lim_{k\to\infty} r_k(t)$ are well-defined, and

$$\sup_{t\geq 0}|(s_k(t), i_k(t), r_k(t)) - (s(t), i(t), r(t))| \leq (1-\rho)^k.$$

Next, we will show that $S^{(\rho)}_{n,r}(t)$ is local, meaning that it converges to $s_r(t)$ uniformly in t. This step is essential since $S^{(\rho)}_{n,r}$ is not a continuous function of t with n discontinuities at the time of infection of each node. The proof is based on conditioning on the structure of the graph in the local neighborhood and then using the tightness of graphs with local limits to bound the probability.

Lemma 5 **(Convergence of Local Approximation)**. *Let $(G_n)_{n\geq 1}$ be a graph sequence that converges locally in probability to $(G, o) \sim \mu$. Then for any $t \in [0, \infty]$, $\mathbb{E}_\Xi[S^{(\rho)}_{n,r}(t)] \xrightarrow{\mathbb{P}} s_r(t)$, where the convergence in probability is with respect to the randomness of G_n.*

Now, to prove Theorem 3, first we can apply Lemmas 1 and 3 to prove that $S^{(\rho)}_{n,r}(t)$ is a good approximation of $S^{(\rho)}_n(t)$. Then we can subsequently apply Lemma 4 and 5 to prove convergence of $S^{(\rho)}_n(t)$ to $s(t)$ uniformly in t. Finally, Theorem 4 is a corollary of Theorems 2 and 3, as convergent graphs satisfy both tightness and stable neighborhood conditions. The details appear in Appendix C.

4.4. Generalizations

4.4.1. Proofs for General Epidemics We start with Corollary 1, which follows very similar steps as in the proof of Theorems 3 and 4, with a small subtlety. Before, it was enough to prove the concentrations for the number of susceptible nodes, and that naturally led to concentration for recovered and, subsequently, infectious nodes. Now, we need to stretch this idea a bit. We will show that the number of nodes in states S, or $\mathcal{D}_1$ through $\mathcal{D}_i$ are concentrated. Then, the proof works similarly to before; effectively, you can think of nodes in the union of these $i+1$ states as a new susceptible state. Once we establish the convergence for these combined states, the convergence of nodes in a specific state $\mathcal{D}_i$ is proved by subtracting from these larger unions. See Appendix D.1.

4.4.2. Proofs for General Starting Configuration Again, the main idea is to truncate the epidemic to a constant radius r and argue that this offers a good approximation of the epidemic on the entire graph. We can generalize the first-moment Lemma 4 to the case that the initial condition is locally reachable. Further, the bound on the second moment is a direct implication of Lemma 3, since the truncated epidemic of two nodes at a large distance are independent. The only caveat is that we need to add the distance ℓ, which determines the dependency radius of the starting configuration P_ℓ. See Appendix D.2.

5. Empirical Validation

To empirically validate the theoretical results, we carried out experiments using both synthetic and real-world networks. The goal is to assess the applicability and accuracy of Algorithm 1 in predicting epidemics on synthetic and real-world networks.

Synthetic Networks: We generated synthetic networks using two well-established models: Preferential Attachment (Barabási and Albert 1999) and Random Geometric Graphs (Gilbert 1959). For both models, we ensured an average degree close to 6 and created synthetic networks of varying sizes, ranging from 500 to $10,000$ nodes. Detailed generation procedures are provided in Appendix G.

Copenhagen Interaction Network: To further validate the estimator's effectiveness in a real-world setting, we used network data from the Copenhagen Networks Study (Sapiezynski et al. 2019). This dataset records interactions among university students at 5-minute intervals over a four-week duration. The interactions were based on Received Signal Strength Indicator (RSSI) values from Bluetooth signal strength measurements, reflecting the physical proximity between individuals. We focused on a specific 12-hour interval (8 am to 8 pm) on the fourth day of the study, considering only those connections within a 6-foot range (RSSI value of -74.25). The resulting graph has 422 nodes with an average degree of 7.89 (see Figure 6).

SafeGraph San Francisco Network: This dataset is derived from SafeGraph mobility data (Chang et al. 2021) for San Francisco County. The data is structured as a bipartite network with time-varying edges between Census Block Groups (CBG) — units of 600 to 3,000 people — and Points of Interest (POI). Edge weights represent the number of CBG visitors to a POI in a given hour.

To construct our edge-weighted network, we aggregated the mobility data for San Francisco County over six hours on March 1, 2020 (6 am to 12 pm). The resulting network comprises $28,713$ POIs and $2,943$ CBGs, for a total of $31,656$ nodes and $82,022$ weighted edges (see Figure 7). In our analysis, we treated each POI and CBG as individual nodes, with edge weights used as the transmission rates.

Experiment Details and Epidemic Parameters: In our experiments, the infection and recovery times were drawn from exponential random variables with rates set to 1. For datasets with edge weights, these weights scale the infection rate. The initial infection probability for a node was set to 0.01. For the estimator (3), we used a total of 10 queries ($q = 10$) and varied the input budget (k) in Algorithm 1 from 2 to 9. In our implementation of Algorithm 1, if the backward process probes k nodes without encountering any initially infected node, we randomly select one of the furthest nodes (in terms of graph distance from the root) to be initially infected[3]. We repeated the experiment $1,000$ times to compute confidence intervals[4].

[3] This modification–introduced to accelerate convergence in practice—does not affect the theoretical guarantees since, for sufficiently large k, the probability of not encountering any initially infected node is very low.

[4] See the code base in Alimohammadi et al. (2025).

Performance Evaluation: We compare the estimtor's output to the ground-truth SIR process (obtained via 1,000 iterations using the Epidemic on Network package in Python (Miller and Ting 2019, Kiss et al. 2017)). The average of these 1,000 runs was considered the ground-truth time evolution. See Figures 2 to 5. Throughout our assessment, distinct scenarios unfolded. For the Copenhagen dataset and the Preferential Attachment Model, $k = 6$ per query was optimal, while for San Francisco and Geometric Random Graphs, a budget of $k = 9$ yielded the best performance. With $q = 10$ queries, this corresponds to sampling only 60 nodes in the former and 90 nodes in the latter cases. Remarkably, this translated to using only 0.28% of San Francisco, 14.9% of Copenhagen, 0.9% of Random Geometric Graphs, and 0.6% of Preferential Attachment nodes, yet maintaining high prediction accuracy.We further assessed performance of the estimated time series via Euclidean distance and Pearson correlation with ground-truth (see Tables 1 to 4), and examined the impact of graph size on estimator accuracy (Table 5). Additional results on robustness to mis-specified distributions and weighted sampling (see Theorems 6 and 5) are in Appendix G.

6. Conclusion

In this work, we developed a new approach for forecasting epidemic dynamics on networks. By introducing a local estimator with robust theoretical guarantees, we demonstrate that epidemic behavior is fundamentally *local*. Notably, our empirical results validate the predictive power of our estimator across datasets.

Our results have important policy implications. As the world faces recurring epidemics, our findings highlight the advantages of targeted data-gathering strategies. Instead of attempting to collect complete network data, policymakers and researchers may consider relying on local network structures, which ensures both efficiency in collection and predictive performance. This stands in contrast to traditional mean-field or full-network simulation models (Birge et al. 2022, Acemoglu et al. 2023), which require either strong mixing assumptions or vast amounts of data.

In our work, we also integrate the theory of local convergence into operations research. Given the vast range of problems governed by network dynamics, numerous applications benefit from this theory. The diffusion of misinformation (Mostagir and Siderius 2023), the complexities of viral marketing (Ho et al. 2002, Manshadi et al. 2020), pricing for accelerating diffusion (Kalish and Lilien 1983, Shen et al. 2011), network learning (Hu et al. 2019), and many other studies have traditionally been restricted by assumptions about network models. However, our method offers a fresh lens, providing a deeper understanding of network dynamics without necessitating network model assumptions.

Acknowledgments

We gratefully acknowledge the constructive comments and suggestions provided by the anonymous reviewers, which have greatly improved this manuscript. This research was partially conducted during the visit of all authors to the Simons Institute of Theoretical Computing in the "Graph Limits and Processes on Networks: From Epidemics to Misinformation" program in the fall of 2022. The work of RvdH is supported in parts by the NWO through the Gravitation NETWORKS grant 024.002.003.

References

Acemoglu D, Makhdoumi A, Malekian A, Ozdaglar A (2023) Testing, voluntary social distancing, and the spread of an infection. *Operations Research* .

Ajorlou A, Jadbabaie A, Kakhbod A (2018) Dynamic pricing in social networks: The word-of-mouth effect. *Management Science* 64(2):971–979.

Akbarpour M, Malladi S, Saberi A (2018) Diffusion, seeding, and the value of network information. *Proceedings of the 2018 ACM Conference on Economics and Computation*, 641–641.

Aldous D, Steele J (2004) The objective method: probabilistic combinatorial optimization and local weak convergence. *Probability on discrete structures*, volume 110 of *Encyclopaedia Math. Sci.*, 1–72 (Springer, Berlin), URL `http://dx.doi.org/10.1007/978-3-662-09444-0_1`.

Alimohammadi Y, Borgs C, Saberi A (2022) Algorithms using local graph features to predict epidemics. *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 3430–3451 (SIAM).

Alimohammadi Y, Borgs C, Saberi A (2023) Locality of random digraphs on expanders. *The Annals of Probability* 51(4):1249–1297.

Alimohammadi Y, Borgs C, van der Hofstad R, Saberi A (2025) localsir: A code base for local sir epidemic simulations. `https://github.com/yalimohammadi/localSIR`, accessed: 2025-03-19.

Amini H, Minca A (2016) Inhomogeneous financial networks and contagious links. *Operations Research* 64(5):1109–1120.

Andreasen V, Lin J, Levin SA (1997) The dynamics of cocirculating influenza strains conferring partial cross-immunity. *Journal of Mathematical Biology* 35:825–842.

Aparicio JP, Capurro AF, Castillo-Chavez C (2000) Transmission and dynamics of tuberculosis on generalized households. *Journal of Theoretical Biology* 206(3):327–341.

Baek J, Farias VF, Georgescu A, Levi R, Peng T, Sinha D, Wilde J, Zheng A (2021) The limits to learning a diffusion model. *Proceedings of the 22nd ACM Conference on Economics and Computation*, 130–131.

Bajardi P, Poletto C, Ramasco JJ, Tizzoni M, Colizza V, Vespignani A (2011) Human mobility networks, travel restrictions, and the global spread of 2009 h1n1 pandemic. *PloS one* 6(1):e16591.

Ball F, Britton T (2022) Epidemics on networks with preventive rewiring. *Random Structures & Algorithms* 61(2):250–297.

Ball F, Sirl D, Trapman P (2010) Analysis of a stochastic sir epidemic on a random network incorporating household structure. *Mathematical Biosciences* 224(2):53–73.

Bampo M, Ewing MT, Mather DR, Stewart D, Wallace M (2008) The effects of the social structure of digital networks on viral marketing performance. *Information Systems Research* 19(3):273–290.

Banerjee A, Chandrasekhar AG, Duflo E, Jackson MO (2013) The diffusion of microfinance. *Science* 341(6144):1236498.

Barabási AL, Albert R (1999) Emergence of scaling in random networks. *science* 286(5439):509–512.

Bartlett M (1949) Some evolutionary stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)* 11(2):211–229.

Bass FM (1969) A new product growth for model consumer durables. *Management science* 15(5):215–227.

Bastani H, Drakopoulos K, Gupta V, Vlachogiannis I, Hadjichristodoulou C, Lagiou P, Magiorkinis G, Paraskevis D, Tsiodras S (2021) Efficient and targeted COVID-19 border testing via reinforcement learning. *Nature* 599(7883):108–113.

Beaudry NJ, Renner R (2012) An intuitive proof of the data processing inequality. *Quantum Information & Computation* 12(5-6):432–441.

Benjamini I, Schramm O (2001) Recurrence of distributional limits of finite planar graphs. *Electron. J. Probab.* **6**:no. 23, 13 pp. (electronic), ISSN 1083-6489, URL `http://dx.doi.org/10.1214/EJP.v6-96`.

Berger N, Borgs C, Chayes J, Saberi A (2014) Asymptotic behavior and distributional limits of preferential attachment graphs. *Annals of Probability* **42**(1):1–40, ISSN 0091-1798, URL `http://dx.doi.org/10.1214/12-AOP755`.

Bernoulli D (1760) Essai d'une nouvelle analyse de la mortalite causee par la petite verole, et des avantages de l'inoculation pour la prevenir. *Histoire de l'Acad., Roy. Sci.(Paris) avec Mem* 1–45.

Bertsimas D, Boussioux L, Cory-Wright R, Delarue A, Digalakis V, Jacquillat A, Kitane DL, Lukin G, Li M, Mingardi L, et al. (2021) From predictions to prescriptions: A data-driven response to COVID-19. *Health Care Management Science* 24:253–272.

Bhamidi S, van der Hofstad R, Komjáthy J (2014) The front of the epidemic spread and first passage percolation. *Journal Of Applied Probability* 51(A):101–121.

Birge JR, Candogan O, Feng Y (2022) Controlling epidemic spread: Reducing economic losses with targeted closures. *Management Science* 68(5):3175–3195.

Blower SM, Mclean AR, Porco TC, Small PM, Hopewell PC, Sanchez MA, Moss AR (1995) The intrinsic transmission dynamics of tuberculosis epidemics. *Nature Medicine* 1(8):815–821.

Bollobás B, Janson S, Riordan O (2007) The phase transition in inhomogeneous random graphs. *Random Structures Algorithms* **31**(1):3–122, ISSN 1042-9832.

Bordenave C (2016) Lecture notes on random graphs and probabilistic combinatorial optimization, version April 8, 2016. Available at `http://www.math.univ-toulouse.fr/~bordenave/coursRG.pdf`.

Britton T, Pardoux E, Ball F, Laredo C, Sirl D, Tran VC (2019) *Stochastic epidemic models with inference*, volume 2255 (Springer).

Budhiraja A, Dupuis P, Fischer M (2012) Large deviation properties of weakly interacting processes via weak convergence methods. *Annals of Probability* 40(1):74–102.

Chang S, Pierson E, Koh PW, Gerardin J, Redbird B, Grusky D, Leskovec J (2021) Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 589(7840):82–87.

Chen J, Hoops S, Marathe A, Mortveit H, Lewis B, Venkatramanan S, Haddadan A, Bhattacharya P, Adiga A, Vullikanti A, et al. (2022) Effective social network-based allocation of COVID-19 vaccines. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4675–4683.

Chin A, Eckles D, Ugander J (2022) Evaluating stochastic seeding strategies in networks. *Management Science* 68(3):1714–1736.

Coppini F, Dietert H, Giacomin G (2020) A law of large numbers and large deviations for interacting diffusions on erdős–rényi graphs. *Stochastics and Dynamics* 20(02):2050010.

Cvjetanović B, Grab B, Uemura K (1971) Epidemiological model of typhoid fever and its use in the planning and evaluation of antityphoid immunization and sanitation programmes. *Bulletin of the World Health Organization* 45(1):53.

Decreusefond L, Dhersin JS, Moyal P, Tran VC (2012) Large graph limit for an sir process in random network with heterogeneous connectivity. *The Annals of Applied Probability* 22(2):541 – 575, URL `http://dx.doi.org/10.1214/11-AAP773`.

Dembo A, Montanari A (2010) Gibbs measures and phase transitions on sparse random graphs. *Braz. J. Probab. Stat.* **24**(2):137–211, ISSN 0103-0752, URL `http://dx.doi.org/10.1214/09-BJPS027`.

Dimitrov NB, Meyers LA (2010) Mathematical approaches to infectious disease prediction and control. *Risk and Optimization in an Uncertain World*, 1–25 (INFORMS).

Drakopoulos K, Zheng F (2017) Network effects in contagion processes: Identification and control. *Columbia Business School Research Paper* (18-8).

Eckles D, Esfandiari H, Mossel E, Rahimian MA (2022) Seeding with costly network information. *Operations Research* (4):2318–2348.

Eubank S, Guclu H, Anil Kumar V, Marathe MV, Srinivasan A, Toroczkai Z, Wang N (2004) Modelling disease outbreaks in realistic urban social networks. *Nature* 429(6988):180–184.

Feder G, Umali DL (1993) The adoption of agricultural innovations: a review. *Technological forecasting and social change* 43(3-4):215–239.

Ford EW, Menachemi N, Phillips MT (2006) Predicting the adoption of electronic health records by physicians: when will health care be paperless? *Journal of the American Medical Informatics Association* 13(1):106–112.

Ganguly A, Ramanan K (2024) Hydrodynamic limits of non-markovian interacting particle systems on sparse graphs. *Electronic Journal of Probability* 29:1–63.

Garavaglia A, Hazra R, van der Hofstad R, Ray R (2022) Universality of the local limit in preferential attachment models, arXiv:2212.05551 [math.PR].

Garavaglia A, van der Hofstad R, Litvak N (2020) Local weak convergence for pagerank. *The Annals of Applied Probability* 30(1):40–79.

Gilbert EN (1959) Random graphs. *The Annals of Mathematical Statistics* 30(4):1141–1144.

Goel S, Anderson A, Hofman J, Watts DJ (2016) The structural virality of online diffusion. *Management Science* 62(1):180–196.

Goldsmith-Pinkham P, Imbens GW (2013) Social networks and the identification of peer effects. *Journal of Business & Economic Statistics* 31(3):253–264.

Graham BS (2008) Identifying social interactions through conditional variance restrictions. *Econometrica* 76(3):643–660.

Gupta S, Hill A, Kwiatkowski D, Greenwood AM, Greenwood BM, Day KP (1994) Parasite virulence and disease patterns in plasmodium falciparum malaria. *Proceedings of the National Academy of Sciences* 91(9):3715–3719.

Gupta S, Starr MK, Farahani RZ, Asgari N (2022) Om forum—pandemics/epidemics: Challenges and opportunities for operations management research. *Manufacturing & Service Operations Management* 24(1):1–23.

Heesterbeek H, Anderson RM, Andreasen V, Bansal S, De Angelis D, Dye C, Eames KT, Edmunds WJ, Frost SD, Funk S, et al. (2015) Modeling infectious disease dynamics in the complex landscape of global health. *Science* 347(6227):aaa4339.

Hethcote HW, Yorke JA (1984) *Gonorrhea transmission dynamics and control*, volume 56 (Springer).

Ho TH, Savin S, Terwiesch C (2002) Managing demand and sales dynamics in new product diffusion under supply constraint. *Management science* 48(2):187–206.

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* 47(260):663–685.

Hu MM, Yang S, Xu DY (2019) Understanding the social learning effect in contagious switching behavior. *Management Science* 65(10):4771–4794.

Isham V (1988) Mathematical modelling of the transmission dynamics of hiv infection and aids: a review. *Journal of the Royal Statistical Society Series A: Statistics in Society* 151(1):5–30.

Jackson MO, Yariv L (2005) Diffusion on social networks. *Economie Publique (Public Economics)* 16(1):3–16.

Jacobsen KA, Burch MG, Tien JH, Rempała GA (2018) The large graph limit of a stochastic epidemic model on a dynamic multilayer network. *Journal of Biological Dynamics* 12(1):746–788.

Janson S, Luczak M, Windridge P (2014) Law of large numbers for the sir epidemic on a random graph with given degrees. *Random Structures & Algorithms* 45(4):726–763.

Kalish S, Lilien GL (1983) Optimal price subsidy policy for accelerating the diffusion of innovation. *Marketing Science* 2(4):407–420.

Kaplan EH (2020) Om forum—COVID-19 scratch models to support local decisions. *Manufacturing & Service Operations Management* 22(4):645–655.

Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 137–146.

Kermack WO, McKendrick AG (1927) 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.

Kim L, Abramson M, Drakopoulos K, Kolitz S, Ozdaglar A (2014) Estimating social network structure and propagation dynamics for an infectious disease. *Social Computing, Behavioral-Cultural Modeling and Prediction: 7th International Conference, SBP 2014, Washington, DC, USA, April 1-4, 2014. Proceedings 7*, 85–93 (Springer).

Kiss IZ, Miller JC, Simon PL, et al. (2017) Mathematics of epidemics on networks. *Cham: Springer* 598.

Kojaku S, Hébert-Dufresne L, Mones E, Lehmann S, Ahn YY (2021) The effectiveness of backward contact tracing in networks. *Nature physics* 17(5):652–658.

Komjáthy J, Lodewijks B (2020) Explosion in weighted hyperbolic random graphs and geometric inhomogeneous random graphs. *Stochastic Process. Appl.* **130**(3):1309–1367, ISSN 0304-4149, URL `http://dx.doi.org/10.1016/j.spa.2019.04.014`.

Krioukov D, Papadopoulos F, Kitsak M, Vahdat A, Boguñá M (2010) Hyperbolic geometry of complex networks. *Phys. Rev. E (3)* 82(3):036106, 18, ISSN 1539-3755, URL `https://doi-org.dianus.libr.tue.nl/10.1103/PhysRevE.82.036106`.

Kurauskas V (2022) On local weak limit and subgraph counts for sparse random graphs. *Journal of Applied Probability* 59(3):755–776, URL `http://dx.doi.org/10.1017/jpr.2021.84`.

Lacker D, Ramanan K, Wu R (2023) Local weak convergence for sparse networks of interacting processes. *The Annals of Applied Probability* 33(2):843–888.

Lala SG, Little KM, Tshabangu N, Moore DP, Msandiwa R, Van Der Watt M, Chaisson RE, Martinson NA (2015) Integrated source case investigation for tuberculosis (tb) and hiv in the caregivers and household contacts of hospitalised young children diagnosed with tb in south africa: an observational study. *PLoS One* 10(9):e0137518.

Larson RC (2007) Simple models of influenza progression within a heterogeneous population. *Operations research* 55(3):399–412.

Lashari AA, Serafimović A, Trapman P (2021) The duration of a supercritical sir epidemic on a configuration model. *Electronic Journal of Probability* 26:1–49.

Lee YJ, Hosanagar K, Tan Y (2015) Do i follow my friends or the crowd? information cascades in online movie ratings. *Management Science* 61(9):2241–2258.

Lloyd-Smith JO, George D, Pepin KM, Pitzer VE, Pulliam JR, Dobson AP, Hudson PJ, Grenfell BT (2009) Epidemic dynamics at the human-animal interface. *science* 326(5958):1362–1367.

Lobel I, Sadler E, Varshney LR (2017) Customer referral incentives and social media. *Management Science* 63(10):3514–3529.

Mamani H, Chick SE, Simchi-Levi D (2013) A game-theoretic model of international influenza vaccination coordination. *Management Science* 59(7):1650–1670.

Mandal S, Sarkar RR, Sinha S (2011) Mathematical models of malaria-a review. *Malaria Journal* 10(1):1–19.

Manshadi V, Misra S, Rodilitz S (2020) Diffusion in random networks: Impact of degree distribution. *Operations research* (6):1722–1741.

May RM, Anderson RM (1987) Commentary transmission dynamics of hiv infection. *Nature* 326(137):10–1038.

Mbivnjo EL, Lynch M, Huws JC (2022) Measles outbreak investigation process in low-and middle-income countries: a systematic review of the methods and costs of contact tracing. *Journal of Public Health* 30(10):2407–2426.

Mihara S, Tsugawa S, Ohsaki H (2015) Influence maximization problem for unknown social networks. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 1539–1546.

Milewska M, van der Hofstad R, Zwart B (2025) Sir on locally converging dynamic random graphs. *arXiv preprint arXiv:2501.09623* .

Miller JC, Ting T (2019) Eon (epidemics on networks): a fast, flexible python package for simulation, analytic approximation, and analysis of epidemics on networks. *Journal of Open Source Software* 4(44):1731.

Mostagir M, Siderius J (2023) Social inequality and the spread of misinformation. *Management Science* 69(2):968–995.

Mukherjee UK, Seshadri S (2022) Epidemic modeling, prediction, and control. *Tutorials in Operations Research: Emerging and Impactful Topics in Operations*, 1–35 (INFORMS).

Müller J, Kretzschmar M (2021) Forward thinking on backward tracing. *Nature Physics* 17(5):555–556.

Netrapalli P, Sanghavi S (2012) Learning the graph of epidemic cascades. *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, 211–222, SIGMETRICS '12 (New York, NY, USA: Association for Computing Machinery), ISBN 9781450310970, URL `http://dx.doi.org/10.1145/2254756.2254783`.

Raymenants J, Geenen C, Thibaut J, Nelissen K, Gorissen S, Andre E (2022) Empirical evidence on the efficiency of backward contact tracing in covid-19. *Nature Communications* 13(1):4750.

Ross R, Hudson HP (1917) An application of the theory of probabilities to the study of a priori pathometry.—part iii. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character* 93(650):225–240.

Sapiezynski P, Stopczynski A, Lassen DD, Lehmann S (2019) Interaction data from the copenhagen networks study. *Scientific Data* 6(1):315.

Scarpino SV, Petri G (2019) On the predictability of infectious disease outbreaks. *Nature Communications* 10(1):898.

Shen W, Duenyas I, Kapuscinski R (2011) New product diffusion decisions under supply constraints. *Management Science* 57(10):1802–1810.

Stein S, Eshghi S, Maghsudi S, Tassiulas L, Bellamy RK, Jennings NR (2017) Heuristic algorithms for influence maximization in partially observable social networks. *SocInf@ IJCAI*, 20–32.

Trapman P (2007) On analytical approaches to epidemics on networks. *Theoretical population biology* 71(2):160–173.

van der Hofstad R (2024) Random graphs and complex networks. Vol. 2, see `http://www.win.tue.nl/∼rhofstad/NotesRGCNII.pdf`.

van der Hofstad R, Hoorn Pvd, Maitra N (2023) Local limits of spatial inhomogeneous random graphs. *Advances in Applied Probability* 1–48, URL `http://dx.doi.org/10.1017/apr.2022.61`.

van der Hofstad R, Komjáthy J, Vadon V (2021) Random intersection graphs with communities. *Adv. in Appl. Probab.* 53(4):1061–1089, ISSN 0001-8678, URL `http://dx.doi.org/10.1017/apr.2021.12`.

van der Hofstad R, van Leeuwaarden JS, Stegehuis C (2016) Hierarchical configuration model. *Internet Mathematics* .

Watts DJ (2002) A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences* 99(9):5766–5771.

Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO (2012) Quantifying the impact of human mobility on malaria. *Science* 338(6104):267–270.

Wilder B, Immorlica N, Rice E, Tambe M (2018) Maximizing influence in an unknown social network. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Wu JT, Wein LM, Perelson AS (2005) Optimization of influenza vaccine selection. *Operations Research* 53(3):456–476.

Yang Y, Nishikawa T, Motter AE (2017) Small vulnerable sets determine large network cascades in power grids. *Science* 358(6365):eaan3184.

Appendix A: Concentration of Epidemic - Proof Details

A.1. Local Approximation - Proof of Bounds on the First Moment

In this section, we delve into two closely related proofs, both of which focus on the concentration of the first moment of a local approximation. These proofs leverage the inherent structure of our models and the local reachability of nodes to shed light on the nuances of approximation within the confines of the given parameters.

The first proof, corresponding to Lemma 1, shows how the path to an initial infection can be bounded within a specific radius due to the fact that initially infected nodes are 'locally reachable'. The second proof (Lemma 4) builds upon the foundation set by the first. By concentrating on the characteristics of the nodes and their infection times, we derive insights into the monotonic behaviors of our models and how they converge in specific scenarios.

Proof of Lemma 1. The proof is a standard argument showing that the shortest path (in terms of $\mathrm{dist}_{(T)}$) from a node to its initial infection can be constrained within a bounded radius due to the fact that the initially infected nodes are 'locally reachable.'

We first draw the marks corresponding to the contact and recovery times, and we only keep the initial infection random. Recall the notations of $T^{(r)}$ which maps a rooted marked graphs $(G, o, \mathcal{M}(G))$ to the infection time of o under the assumption that the network is restricted to $B_r(G, o)$. Using this, $T^{(\infty)}(v)$ refers to the actual infection time of v in G_n. Note that the difference in $S_n(t)$ and $S_{n,r}(t)$ is in the set of nodes for which $T^{(\infty)}(v) \neq T^{(r)}(v)$, i.e.,

$$\mathbb{E}_{\Xi}\left[\sup_{t\geq 0}\left|\mathscr{E}_n^{(\rho)}(t) - \mathscr{E}_{n,r}^{(\rho)}(t)\right|\right] \leq \mathbb{E}_{\Xi}\left[\frac{1}{n}\left|\{v : T^{(\infty)}(v) \neq T^{(r)}(v)\}\right|\right] = \mathbb{P}_{\Xi}\Big(T^{(\infty)}(o_n) \neq T^{(r)}(o_n)\Big),$$

where o_n is a uniformly random chosen node from $V(G_n)$. Consider the shortest path that identifies the infection time $T^{(\infty)}(v)$. If $T^{(\infty)}(v) \neq T^{(r)}(v)$ then the graph length of this shortest path is at least $r+1$, otherwise the infection time was already identified in the r-neighborhood. Moreover, the initial r nodes of this shortest path toward determining $T^{(\infty)}(v)$ should not contain any initially infected node. If it did, the path would reach another initially infected node faster. Therefore,

$$\mathbb{P}\Big(T^{(\infty)}(o_n) \neq T^{(r)}(o_n) \mid \mathcal{M}(G_n)\Big) \leq (1-\rho)^r,$$

where the probability is with respect to the initial infections. Now, if we take the probability over the marks of epidemic, we get that

$$\mathbb{P}_{\Xi}\Big(T^{(\infty)}(o_n) \neq T^{(r)}(o_n)\Big) \leq (1-\rho)^r,$$

which in turn provides an upper bound for $\mathbb{E}_{\Xi}\left[\frac{1}{n}\left|\{v : T^{(\infty)}(v) \neq T^{(r)}(v)\}\right|\right]$. □

Proof of Lemma 4. The first part of the lemma follows from the same argument as in the proof of Lemma 1. Then by noting that $s_k(t)$ is monotone decreasing in k, the limit $s(t) = \lim_{k\to\infty} s_k(t)$ is well-defined, and the bound on their difference follows from the first part. Similarly, $r_k(t)$ is monotone increasing in k, and the same argument holds. Finally, we can finish the argument by noting that $i_k(t) = 1 - s_k(t) - r_k(t)$. □

A.2. Local Approximation - Proof of Bounds on the Second Moment

In this section, we delve into the second moment of our local approximation. In our first proof, Lemma 2, we use the independence of events for nodes that are sufficiently far apart in the graph, allowing us to derive a precise bound for the variance of $S^{(\rho)}_{n,r}(t)$.

Proof of Lemma 2. The proof follows from the fact that the events $\{t < T^{(r)}(x)\}$ and $\{t < T^{(r)}(y)\}$ are independent if $\text{dist}_{G_n}(x,y) > 2r$. Let $Z_v(t) = \mathbb{1}\{t < T^{(r)}(G_n, v, \mathcal{M}(G_n))\}$. Then, we can write

$$\begin{aligned}
n^2\text{Var}(S^{(\rho)}_{n,r}(t)) &= \mathbb{E}\Big[\big(\sum_{v\in V(G_n)} Z_v(t) - \sum_{v\in V(G_n)} \mathbb{E}(Z_v(t))\big)^2\Big] \\
&= \sum_{v\in V(G_n)} \mathbb{E}\Big[\big(Z_v(t) - \mathbb{E}(Z_v(t))\big)^2\Big] + \sum_{\text{dist}_{G_n}(v,u)\leq 2r} \mathbb{E}\Big[\big(Z_v(t) - \mathbb{E}(Z_v(t))\big)\big(Z_u(t) - \mathbb{E}(Z_u(t))\big)\Big] \\
&\quad + \sum_{\text{dist}_{G_n}(v,u)> 2r} \mathbb{E}\Big[\big(Z_v(t) - \mathbb{E}(Z_v(t))\big)\big(Z_u(t) - \mathbb{E}(Z_u(t))\big)\Big] \\
&= \sum_{v\in V(G_n)} \mathbb{E}\Big[\big(Z_v(t) - \mathbb{E}(Z_v(t))\big)^2\Big] + \sum_{\text{dist}_{G_n}(v,u)\leq 2r} \mathbb{E}\Big[\big(Z_v(t) - \mathbb{E}(Z_v(t))\big)\big(Z_u(t) - \mathbb{E}(Z_u(t))\big)\Big].
\end{aligned}$$

Here, we use the fact that if $\text{dist}_{G_n}(u,v) \geq 2r$ then $Z_u(t)$ and $Z_v(t)$ are independent. Therefore,

$$\mathbb{E}\Big[\big(Z_v(t) - \mathbb{E}(Z_v(t))\big)\big(Z_u(t) - \mathbb{E}(Z_u(t))\big)\Big] = 0.$$

To finish the proof note that $0 \leq Z_v(t) \leq 1$, so an obvious upper bound is $|Z_v(t) - \mathbb{E}(Z_v(t))| \leq 1$. Therefore,

$$n^2\text{Var}(S^{(\rho)}_{n,r}(t)) \leq n + n^2\varepsilon_{2r}(G_n).$$

Note that all the bounds are independent of t, which finishes the proof. □

Subsequently, we extend this analysis, factoring in the randomness of graph G_n. The following proof emphasizes the stability of local neighborhoods and their role in shaping the variance.

Proof of Lemma 3. We first write

$$\mathrm{Var}(S_{n,r}^{(\rho)}(t)) = \mathbb{E}(\mathrm{Var}_{\Xi}(S_{n,r}^{(\rho)}(t)|G_n)) + \mathrm{Var}(\mathbb{E}_{\Xi}(S_{n,r}^{(\rho)}(t)|G_n)).$$

The first term can be bounded using Lemma 2. For the second term, we use a standard argument we condition over different graph structures of radius r that appear in the r-neighborhood of a uniform random node. Let $\mathcal{H}$ be the set of all graph structures of $B_r(G_n, v_i)$ for $v_i \in V(G_n)$ up to isomorphism. Also for $H^\star \in \mathcal{H}$, recall that $P_r^{(G_n)}(H^\star) = \frac{1}{|V(G_n)|}\sum_{v\in V(G_n)} \mathbb{1}\{B_r(G_n,v) \simeq H^\star\}$ is the probability that the r-neighborhood of a uniform random node in G_n is isomorphic to $H^\star$, and that $p_r^{(n)}(H^\star) = \mathbb{E}[P_r^{(G_n)}(H^\star)]$ is its expectation with respect to the randomness of G_n. Then

$$\mathbb{E}_{\Xi}(S_{n,r}^{(\rho)}(t)|G_n) = \sum_{H^\star\in\mathcal{H}} P_r^{(G_n)}(H^\star)\mathbb{P}_{\Xi}(t < T^{(r)}(H^\star)|H^\star). \tag{14}$$

To bound the $\mathcal{H}$, we use the tightness argument. Let $\mathcal{H}_k$ be a subset of $\mathcal{H}$ where each graph has a size of at most k. Then

$$\mathbb{E}(S_{n,r}^{(\rho)}(t)|G_n) \le \varepsilon_r(G_n,k) + \sum_{H^\star\in\mathcal{H}_k} P_r^{(G_n)}(H^\star)\mathbb{P}_{\Xi}(t < T^{(r)}(H^\star)|H^\star). \tag{15}$$

Then using tightness, given any $\delta > 0$ for large enough k and n, $\mathbb{P}(\varepsilon_r(G_n,k) \ge \delta) \le 1-\delta$. Therefore, we can assume that $\mathrm{Var}(\varepsilon_r(G_n,k)) \le 2\delta^2$. Also, using the fact that $S_{n,r}(t) \le 1$, we get $\mathrm{Cov}(\varepsilon_r(G_n,k), \mathbb{E}(S_{n,r}^{(\rho)}(t)|G_n)) \le 2\delta$. Therefore,

$$\begin{aligned}
\mathrm{Var}\mathbb{E}(S_{n,r}^{(\rho)}(t)|G_n) &\le 2\delta^2 + 2\delta + \mathbb{E}\Big[\Big(\sum_{H^\star:|H^\star|\le k} \mathbb{P}_{\Xi}(t<T^{(r)}(H^\star)|H^\star)(P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star))\Big)^2\Big]\\
&= 2\delta^2 + 2\delta + \mathbb{E}\Big[\sum_{H^\star:|H^\star|\le k} \mathbb{P}_{\Xi}(t<T^{(r)}(H^\star)|H^\star)(P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star))^2\Big]\\
&\quad + 2\sum_{H^\star:|H^\star|\le k}\sum_{H^{\star\prime}:|H^{\star\prime}|\le k}\Big(\mathbb{P}_{\Xi}(t<T^{(r)}(H^\star)|H^\star)\mathbb{P}_{\Xi}(t<T^{(r)}(H^{\star\prime})|H^{\star\prime})\\
&\qquad \mathbb{E}\Big[(P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star))(P_r^{(G_n)}(H^{\star\prime}) - p_r^{(n)}(H^{\star\prime}))\Big]\Big)\\
&\le 2\delta^2 + 2\delta + \mathbb{E}\Big[\sum_{H^\star}(P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star))^2\Big]\\
&\quad + 2\mathbb{E}\Big[\sum_{H^\star:|H^\star|\le k}\sum_{H^{\star\prime}:|H^{\star\prime}|\le k}|(P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star))(P_r^{(G_n)}(H^{\star\prime}) - p_r^{(n)}(H^{\star\prime}))|\Big].
\end{aligned}$$

By our stable local neighborhood condition in Definition 2, we can bound the last inequality. Let $N_{r,k}$ be the number of rooted graphs of radius r and size k. For a given $\delta' \le \frac{\delta}{N_{r,k}}$, there exists N such that for all $n > N$, and for all $H^\star$, $\mathbb{E}[(P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star))^2] \le \delta'$. Therefore,

$$\mathrm{Var}\Big(\mathbb{E}(S_{n,r}^{(\rho)}(t)|G_n)\Big) \le 3\delta + 4\delta^2. \tag{16}$$

Putting this together with Lemma 2, we get

$$\mathrm{Var}(S_{n,r}^{(\rho)}(t)) \le 3\delta + 4\delta^2 + \mathbb{E}[\varepsilon_{2r}(G_n))] + \frac{1}{n}.$$

Finally, the result follows by applying (18) along with the tightness condition to bound $\varepsilon_{2r}(G_n))$. □

A.3. Proof of Theorem 1 - Concentration of Epidemic for Deterministic Graphs

For simplicity, we state the proof for the number of susceptible nodes, but all the steps are replicable for infectious and recovered nodes. The first step is to prove that $S^{(\rho)}_{n,r}(t)$ concentrates around the time evolution of epidemic $S^{(\rho)}_n(t)$. In particular, the implications of Lemma 2 combined with the Chebyshev inequality shows that for any $\delta' > 0$ and any $t \in [0, \infty]$,

$$\mathbb{P}_\Xi\Big(|S^{(\rho)}_{n,r}(t) - \mathbb{E}_\Xi[S^{(\rho)}_{n,r}(t)]| \geq \delta'\Big) \leq \frac{1}{(\delta')^2}\Big(\frac{1}{n} + \varepsilon_{2r}(G_n)\Big).$$

Building upon this, by employing Lemma 1 and the Markov inequality, we can further deduce that for any $t \in [0, \infty]$,

$$\mathbb{P}_\Xi\Big(|S^{(\rho)}_{n,r}(t) - S^{(\rho)}_n(t)| \geq \delta'\Big) \leq \frac{(1-\rho)^r}{\delta'}. \tag{17}$$

Further, Lemma 1 also implies that

$$\sup_{t\geq 0}\big|\mathbb{E}_\Xi[S^{(\rho)}_{n,r}(t)] - \mathbb{E}_\Xi[S^{(\rho)}_n(t)]\big| \leq \mathbb{E}_\Xi\left[\sup_{t\geq 0}\big|S^{(\rho)}_n(t) - S^{(\rho)}_{n,r}(t)\big|\right] \leq (1-\rho)^r.$$

Combining the previous three inequalities, we get that for any $t \in [0, \infty]$,

$$\mathbb{P}_\Xi\Big(|S^{(\rho)}_n(t) - \mathbb{E}_\Xi[S^{(\rho)}_n(t)]| \geq 2\delta' + (1-\rho)^r\Big) \leq \frac{1}{(\delta')^2}\Big(\frac{1}{n} + \varepsilon_{2r}(G_n)\Big) + \frac{(1-\rho)^r}{\delta'}.$$

Now, to finish the proof of the first part, we need to relate the radius discovered by the algorithm to the number of nodes visited. Let $|B_r(G_n, v)|$ be the number of nodes at distance at most r from v. Recall that $n^2\varepsilon_r(G_n)$ is the number of pairs (u, v) such that $\mathrm{dist}_{G_n}(u, v) \leq r$. Then

$$\begin{aligned}
n^2\varepsilon_r(G_n) &= \sum_v |B_r(G_n, v)| \\
&= \sum_{v:|B_r(G_n,v)|\geq k} |B_r(G_n, v)| + \sum_{v:|B_r(G_n,v)|<k} |B_r(G_n, v)| \\
&\leq n^2\varepsilon_r(G_n, k) + kn(1 - \varepsilon_r(G_n, k)).
\end{aligned}$$

Here in the last inequality, we use the fact that there are $n\varepsilon_r(G_n, k)$ nodes with $|B_r(G_n, v)| \geq k$. We use the obvious bound of $|B_r(G_n, v)| \leq n$ for those, and for the rest of the nodes, we use the bound of $|B_r(G_n, v)| \leq k$. Therefore,

$$\varepsilon_r(G_n) \leq \varepsilon_r(G_n, k) + \frac{k}{n}. \tag{18}$$

Now it remains to bound the deviation of our estimator with q queries, $\hat{S}^{(\rho)}_{q,r,n}(t)$, with $S^{(\rho)}_{n,r}(t)$. We claim that, for any $t \in [0, \infty]$,

$$\mathbb{P}_\Xi\Big(|S^{(\rho)}_{n,r}(t) - \hat{S}^{(\rho)}_{q,r,n}(t)| \geq \delta\Big) \leq 2e^{-2q\delta^2} + \frac{16}{\delta^2}\big(\frac{1}{n} + \varepsilon_{2r}(G_n)\big) \tag{19}$$

Similar as before, define $Z_v(t) = \mathbb{1}\{t < T^{(r)}(G_n, v, \mathcal{M}(G_n))\}$.

Now, let $u_1, u_2, \dots, u_q$ be the set of initial nodes sampled independently by the algorithm. Then

$$S_{n,r}^{(\rho)}(t) = \frac{1}{n} \sum_{v \in V(G_n)} Z_v(t), \qquad \text{and} \qquad \hat{S}_{q,r,n}^{(\rho)}(t) = \frac{1}{q} \sum_{i=1}^{q} Z_{u_i}(t).$$

We couple the marks drawn in the algorithm with the marks of G_n determining $S_{n,r}^{(\rho)}$. Then

$$\mathbb{P}(Z_{u_i}(t) = 1) = S_{n,r}^{(\rho)}(t),$$

and the sampling of u_i is with replacement. So, $Z_{u_i}(t)$ are independent given $\mathcal{M}(G_n)$. Therefore, using Hoeffding inequality, for $t \in [0, \infty]$

$$\mathbb{P}\Big(|S_{n,r}^{(\rho)}(t) - \hat{S}_{q,r,n}^{(\rho)}(t)| \geq \delta \mid \mathcal{M}(G_n)\Big) \leq 2e^{-2q\delta^2}, \tag{20}$$

where the probability is only over the randomness of the starting points of the algorithm u_i. Then, to conclude (19), we can use the variance bound in Lemma 2 to decouple the marks of the epidemic with the algorithm. To formalize this, we use the notation $\mathbb{E}_{alg}$ to show expectation over the randomness of the marks drawn by the algorithm, and as before, we use $\mathbb{E}_{G_n}$ for the randomness of G_n and its marks $\mathcal{M}(G_n)$. Also, with the abuse of notation, we use G_n and alg as input of the local approximation $S_{n,r}^{(\rho)}$ and the estimator $\hat{S}_{q,r,n}^{(\rho)}$ to show the corresponding marks. Our goal is to prove the following lower bound,

$$\mathbb{P}_{\Xi,alg}\Big(|S_{n,r}^{(\rho)}(t, G_n) - \hat{S}_{q,r,n}^{(\rho)}(t, alg)| \leq \delta\Big) \geq 1 - 2e^{-q\delta^2/2} - \frac{16}{\delta^2}\Big(\frac{1}{n} + \varepsilon_{2r}(G_n)\Big). \tag{21}$$

For this purpose, define $E_t = |S_{n,r}^{(\rho)}(t, G_n) - S_{n,r}^{(\rho)}(t, alg)|$ and $\hat{E}_t = |S_{n,r}^{(\rho)}(t, alg) - \hat{S}_{q,r,n}^{(\rho)}(t, alg)|$. Then,

$$\begin{aligned}
&\mathbb{P}_{\Xi,alg}\Big(|S_{n,r}^{(\rho)}(t, G_n) - \hat{S}_{q,r,n}^{(\rho)}(t, alg)| \leq \delta\Big) \geq \\
&\qquad \mathbb{P}_{\Xi,alg}\Big(|S_{n,r}^{(\rho)}(t, G_n) - \hat{S}_{q,r,n}^{(\rho)}(t, alg)| \leq \delta \mid \hat{E}_t \leq \delta/2\Big)\mathbb{P}_{alg}\Big(\hat{E}_t \leq \delta/2\Big) \geq \\
&\qquad \mathbb{P}_{\Xi,alg}\Big(E_t + \hat{E}_t \leq \delta \mid \hat{E}_t \leq \delta/2\Big)\mathbb{P}_{alg}\Big(\hat{E}_t \leq \delta/2\Big),
\end{aligned}$$

where the last bound is using the triangle inequality. As a result,

$$\mathbb{P}_{\Xi,alg}\Big(|S_{n,r}^{(\rho)}(t, G_n) - \hat{S}_{q,r,n}^{(\rho)}(t, alg)| \leq \delta\Big) \geq \mathbb{P}_{\Xi,alg}\Big(E_t \leq \delta/2\Big)\mathbb{P}_{alg}\Big(\hat{E}_t \leq \delta/2\Big).$$

Now, we can apply (20) to bound the coupling between the algorithm and the epidemic, given that the marks are the same.

$$\mathbb{P}_{\Xi,alg}\Big(|S_{n,r}^{(\rho)}(t, G_n) - \hat{S}_{q,r,n}^{(\rho)}(t, alg)| \leq \delta\Big) \geq \big(1 - 2e^{-q\delta^2/2}\big)\mathbb{P}_{\Xi,alg}\Big(E_t \leq \delta/2\Big).$$

Then we use the variance bound in Lemma 2 for the second event:

$$\mathbb{P}_{\Xi,alg}\Big(|S^{(\rho)}_{n,r}(t,G_n) - \hat{S}^{(\rho)}_{q,r,n}(t,alg)| \leq \delta\Big) \geq \big(1 - 2e^{-q\delta^2/2}\big)\big(1 - \frac{16}{\delta^2}(\frac{1}{n} + \varepsilon_{2r}(G_n))\big),$$

which proves (19). Combining this with (18), we get,

$$\mathbb{P}_{\Xi,alg}\Big(|S^{(\rho)}_{n}(t,G_n) - \hat{S}^{(\rho)}_{q,r,n}(t,alg)| \geq \delta\Big) \leq 1 - \big(1 - 2e^{-q\delta^2/2}\big)\big(1 - \frac{16}{\delta^2}(\frac{1+k}{n} + \varepsilon_{r}(G_n,k))\big) + (1-\rho)^r$$

Finally, note that for nodes that $B_r(G_n,v) \leq k$ the algorithm's estimator would match the results given by Z_v. Therefore,

$$\mathbb{E}\Big[\hat{S}^{(\rho)}_{q,k,n}(t) - \hat{S}^{(\rho)}_{q,r,n}(t)\Big] \leq \varepsilon_r(G_n,k)$$

Now using the fact that $\mathbb{E}\Big[\hat{S}^{(\rho)}_{q,k,n}(t) - S^{(\rho)}_{n}(t)\Big] > 0$ we get the desired result. □

A.4. Concentration of Epidemics for Tight Graphs – Proof of Theorem 7.

This is a direct application of Theorem 1 along with the definition of the tightness. □

A.5. Examples on Necessity of our Conditions

Two examples are discussed in this section, highlighting the nuanced impact of specific conditions on epidemic estimations. The first illustrates a scenario where the graph does not satisfy the tightness condition and shows how the time evolution of the epidemic does not concentrate in this case. The second example shows where, with a strictly local starting condition, the final size of the epidemic does not concentrate around its mean even if the underlying network is tight.

Example 3 (Necessity of the tightness condition). The necessity of the condition in Definition 1 for local estimation of epidemics becomes apparent when examining specific graph structures. Consider the star graph, wherein a central node connects to $n-1$ peripheral nodes. In this scenario, the final infection size and the time evolution of the epidemic fail to concentrate. To see this, note that except for the ρ fraction of peripheral nodes that are initially infected, the rest of them can only get infected through the central node. Due to its high degree, the central node will, with a high probability, eventually become infected. Assuming the recovery rate to be equal to the transmission rate, the number of nodes to which the central node transmits the disease becomes a uniform random variable within the range of 0 to $(1-\rho)n$. Consequently, without observing the recovery time of the central node, estimating the final infection size or the time evolution becomes infeasible. This example does not satisfy the tightness condition since $|B_2(G_n,v)| = n$ for every node v. Definition 1 controls the influence of large-degree nodes and imposes a regularity on the system, leading to more stable and predictable infection spread across the graph. ◀

Example 4 (Strict locality is not enough for convergence of final size). Consider three distinct graphs, each of size n, where the first is formed by blowing up each node of a 3-regular random graph with a triangle, and the second and third are standard 3-regular random graphs. We add an edge between two random nodes of the first and second graph and two random nodes of the second and third graph. Suppose an initial condition is imposed such that nodes within a triangle are infected while others are susceptible. Under this starting configuration, the first graph becomes entirely infected, while the second and third remain susceptible. The evolution of the epidemic then depends on a single bridging node in the second graph, leading to three potential outcomes for the final infection size: a rapid die-out in the second graph, a linear spread in the second but not the first, or a linear spread in both. Consequently, the final size does not converge to a deterministic value. ◄

Appendix B: Proof of Theorem 2 - Concentration of Epidemic for Random Graphs

The first part of the theorem on the concentration of the epidemic follows the exact same argument as in the proof of Theorem 1. To see it, note that Lemma 3 is enough to give concentration of $S_{n,r}(t)$. To deduce the concentration of $S_n(t)$ from it, we note that Lemma 1 also applies to random graph models (by conditioning on the drawn graph and using the law of total expectation). To deduce the bound on the local estimator, we can follow the same steps, with the change of applying Lemma 3. □

Appendix C: Convergence of Epidemics on Growing Graphs- Proof Details

C.1. Proof of Lemma 5 - Convergence of Local Approximation

The proof is similar to Lemma 3 in the sense that we condition on different structures the r ball of a node can take. Recall equation (14), and that P_Ξ is the probability over the graph marks. Also, recall that $s_r^{(\rho)}(t) = \mu_\Xi\Big(\mathbb{1}\{t < T^{(r)}(G,o)\}\Big)$. As before, we can write

$$s_r^{(\rho)}(t) = \sum_{H^\star \in \mathcal{H}} p_r(H^\star)\mathbb{P}_\Xi\big(t < T^{(r)}(H^\star)\big),$$

where $p_r(H^\star) = \mu(\mathbb{1}\{B_r(G,o) \simeq H^\star\})$ is the probability that the r-neighborhood of the limit graph is isomorphic to $H^\star$. Therefore, the left-hand side of the expression of Lemma 5 can be written as

$$\sup_{t\geq 0}\left|\mathbb{E}_\Xi[S_{n,r}^{(\rho)}(t)] - s_r^{(\rho)}(t)\right| = \sup_{t\geq 0}\left|\sum_{H^\star\in\mathcal{H}} P_\Xi\Big(t < T^{(r)}(H^\star)\Big)\Big(p_r^{(G_n)}(H^\star) - p_r(H^\star)\Big)\right|. \tag{22}$$

For the rest of the proof, we use tightness and stable local neighborhood criteria for graphs converging locally in probability. These conditions are proved in Appendix C.2. Using the tightness condition, we can choose k large enough such that $\mathbb{P}(\varepsilon_{r,k}(G_n) \leq \delta) \geq 1-\delta$. So, as in Lemma 3, if we let $\mathcal{H}_k$ be the set of all locally rooted graphs of size k, then we can approximate the right-hand side of (22) with the sum of $H^\star \in \mathcal{H}_k$. More precisely, using the tightness condition, for any $\delta > 0$, there exists a large enough k, such that

$$\mathbb{P}\Big(\Big|\sup_{t\geq 0}\Big|\sum_{H^\star\in\mathcal{H}} P_\Xi\Big(t < T^{(r)}(H^\star)\Big)\Big(p_r^{(G_n)}(H^\star) - p_r(H^\star)\Big)\Big| - \sup_{t\geq 0}\Big|\sum_{H^\star\in\mathcal{H}_k} P_\Xi\Big(t < T^{(r)}(H^\star)\Big)\Big(p_r^{(G_n)}(H^\star) - p_r(H^\star)\Big)\Big|\Big| \leq \delta\Big) \geq 1-\delta.$$

By applying this to (22), it is enough to prove the following,

$$\sup_{t\geq 0}\Big|\sum_{H^\star\in\mathcal{H}_k} P_\Xi\Big(t < T^{(r)}(H^\star)\Big)\Big(p_r^{(G_n)}(H^\star) - p_r(H^\star)\Big)\Big| \xrightarrow{\mathbb{P}} 0$$

For this purpose, we use the following,

$$\sup_{t\geq 0}\left|\sum_{H^\star\in\mathcal{H}_k} P_\Xi\Big(t < T^{(r)}(H^\star)\Big)\Big(p_r^{(G_n)}(H^\star) - p_r(H^\star)\Big)\right| \leq \sum_{H^\star\in\mathcal{H}_k}\left|p_r^{(G_n)}(H^\star) - p_r(H^\star)\right|.$$

So it remains to provide bound on the right-hand side. This is possible by first observing that $|\mathcal{H}_k|$ is a bounded number since there are finitely many graphs of size k. Second, for each $H^\star$, we can use that

$$\left|p_r^{(G_n)}(H^\star) - p_r(H^\star)\right| \leq \left|p_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star)\right| + \left|p_r^{(n)}(H^\star) - p_r(H^\star)\right|.$$

The first term goes to zero by stable local neighborhood as proved in Appendix C.3. The second term, $\left|p_r^{(n)}(H^\star) - p_r(H^\star)\right| \xrightarrow{\mathbb{P}} 0$ by local convergence in probability (van der Hofstad 2024, Theorem 2.15 (b)). So, the lemma is proved. □

C.2. Proof of Theorem 3 - Convergence of Epidemic

First note that by applying Lemmas 1 and 3 we get that for any $\delta > 0$, and any r and n large enough, and any $t \in [0, \infty]$,

$$\mathbb{P}\Big(|\mathscr{E}_n(t) - \mathbb{E}_\Xi[\mathscr{E}_{n,r}(t)]| \geq \delta\Big) \leq \delta.$$

Then we can subsequently apply Lemma 5 and then Lemma 4 to prove convergence of $S_n(t)$ to $s(t)$ uniformly in t. The proof is similar for infectious and recovered nodes. □

C.3. Proof of Theorem 4 - Local Approximation of the Limit

The tightness condition follows since the distance of two uniform random nodes increases in convergent graphs. More formally, given a sequence of graphs converging in distribution to a limit, and for any given r, $\lim_{n\to\infty}\varepsilon_r(G_n) = 0$, as demonstrated in (van der Hofstad 2024, Corollary 2.20).

Also, the stable neighborhood condition Definition 2 is satisfied by the criterion of local convergence obtained in (van der Hofstad 2024, Theorem 2.15 (b)). To formalize this, note that (van der Hofstad 2024, Theorem 2.15- part b) implies that for any finite rooted graph $H^\star$, and all integers r,

$$P_r^{(G_n)}(H^\star) \xrightarrow{\mathbb{P}} \mu(B_r(G,o) \simeq H^\star).$$

As a result,

$$p_r^{(n)}(H^\star) = \mathbb{E}\big[P_r^{(G_n)}(H^\star)\big] \to \mu(B_r(G,o) \simeq H^\star).$$

Therefore, using a triangle inequality, we get that for any given graph H and integer $r \geq 1$,

$$\begin{aligned}&\mathbb{P}\Big(|P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star)| \geq \delta\Big)\\ &\qquad \leq \mathbb{P}\Big(|P_r^{(G_n)}(H^\star) - \mu(B_r(G,o) \simeq H^\star)| + |\mu(B_r(G,o) \simeq H^\star) - p_r^{(n)}(H^\star)| \geq \delta\Big),\end{aligned}$$

where the right side approaches zero when considering the preceding convergences. Therefore, convergent graphs in probability satisfy the stable local neighborhood condition. As a result, we can apply Theorem 2, establishing that for given any $\delta > 0$, there exists constants N, q_δ, k_δ such that for any $n > N$,

$$\mathbb{P}\Big(|\hat{\mathscr{E}}^{(\rho)}_{n,q_\delta,k_\delta}(t) - \mathscr{E}^{(\rho)}_n(t)| > \delta\Big) \leq \delta. \tag{23}$$

To finish the proof, it is enough to apply Theorem 3, which implies the convergence in probability of $\mathscr{E}^{(\rho)}_n(t)$ to $(s(t), i(t), r(t))$. □

Appendix D: Proof Details for General Epidemics

D.1. Convergence of General Epidemics – Proof of Corollary 1

We follow the proof of Theorem 3 step by step and point out what parts of the proofs need to be changed for the general epidemics. After establishing the conclusion of Theorem 3, the proof of Theorem 4 directly applies in this case.

Recall that the time-varying infectiousness of each node (β_v, τ_v) depends on the ℓ neighborhood of the graph. We start by proving the convergence of the number of susceptible people conditioned on the ℓ neighborhood. As before, we can prove concentration bounds for the number of susceptible nodes and that the truncated epidemics at some $r > \ell$ neighborhood give the right bounds.

Our goal is to prove that for any $\delta > 0$ and any $t \in [0, \infty]$,

$$\lim_{r\to\infty} \lim_{n\to\infty} \mathbb{P}(\big|S^{(\rho)}_n(t) - S^{(\rho)}_{n,r}(t)\big| \geq \delta) = 0. \tag{24}$$

As before, we can define $T^{(r)}(v)$ as the time it takes for node v to leave a susceptible state if the epidemic is confined to its r neighborhood. Then, the exact same argument as in the proof of Lemma 1, for a uniform random vertex $o_n \in V(G_n)$,

$$\mathbb{P}_\Xi\left(T^{(\infty)}(o_n) \neq T^{(r)}(o_n)\right) \leq (1-\rho)^r.$$

To see this, as before, we first fix the marked graph (the nodes, edges, and the distributions β), and keep the epidemic process random. Then consider the shortest path that causes infection time $T^{(\infty)}(v)$ (here the shortest path is with respect to the weights drawn from β_u for each u

on this path). Then, if $T^{(\infty)}(v) \neq T^{(r)}(v)$, the number of vertices in this path would be at least $r+1$. Also, there should not be any initially infected node in the intersection of this path and the r-neighborhood of v, otherwise $T^{(r)}$ would be bounded from above by the weight of this path. So, for each marked graph, the probability of $T^{(\infty)}(v) \neq T^{(r)}(v)$ is bounded by $(1-\rho)^r$, and, hence, if we take the probability over the marks, we prove our claim. As a result,

$$\mathbb{E}_{\Xi}\left[\sup_{t\geq 0} |S_n(t) - S_{n,r}(t)|\right] \leq (1-\rho)^r.$$

A similar first-moment bound works for the number of susceptible in the limit.

Now, we can bound $\mathrm{Var}\Big(S_{n,r}(t)\Big)$ as in Lemma 3. Note that the bounds we used in the second-moment arguments in the proof of Lemma 3 were independent of the specifics of the dynamics of the epidemic, and we only used the fact that $T^{(r)}(v)$ and $T^{(r)}(u)$ of two nodes with $\mathrm{dist}_{G_n}(u,v) > 2r$ are independent. Here, this is true if $\mathrm{dist}_{G_n}(u,v) > 2r + \ell$, where ℓ is added to ensure the transmission probability densities within all nodes of $B_r(G_n, u)$ and $B_r(G_n, v)$ are independent. The two variance and first-moment bounds prove (24).

Next, by applying the proof steps of Lemma 5, we get convergence of $\mathbb{E}_{\Xi}[S_{n,r}(t)]$ to $s_r(t)$, i.e, $\left|\mathbb{E}_{\Xi}[S^{(\rho)}_{n,r}(t)] - s_r(t)\right| \xrightarrow{\mathbb{P}} 0$, where the randomness is with respect to G_n [5] Combining this with the first and second moment results on $S^{(\rho)}_{n,r}(t)$, we get the following convergences for any $t \in [0,\infty]$,

$$S^{(\rho)}_{n}(t) \xrightarrow{\mathbb{P}} s(t).$$

We now extend our previous proof to calculate the proportion of nodes that reside in any of the states $\mathcal{D}_1, \ldots, \mathcal{D}_i$. Let us define $D^{(i)} = \{S, \mathcal{D}_1, \ldots, \mathcal{D}_i\}$ as the combined set of states encompassing S and $\mathcal{D}_1, \ldots, \mathcal{D}_i$. The term $D^{(i)}_n(t)$ represents the proportion of nodes found in any state within $D^{(i)}$ at a given time t in G_n. Analogously, we define $D^{(i)}(t)$ with respect to the limit graph $(G, o) \sim \mu$. Furthermore, as we previously outlined, $T^{(r)}_i(v, G, \mathcal{M}(G))$ as the time v exits state $\mathcal{D}_i$ and enters $\mathcal{D}_{i+1}$, given that the epidemic is truncated at the r-neighborhood. This is fundamentally equivalent to the shortest route from v to the initial infection state plus the sum $t_1 + t_2 + \ldots + t_i$ specific to node v (keeping in mind that t_j denotes the transition period from $\mathcal{D}_j$ to $\mathcal{D}_{j+1}$ as determined by β_v, and after (β_v, τ_v) are sampled, the realizations fo t_is are independent of the neighborhood of v, and the state of epidemics in other nodes). Crucially, considering the given marks, the sole scenario where $T^{(r)}_i(v, G, \mathcal{M}(G))$ differs from $T^{(\infty)}_i(v, G, \mathcal{M}(G))$ is if the shortest route to the initially infected node exceeds a length of r. Otherwise, the contraction time of the disease and $t_1, t_2, \ldots, t_r$ can be coupled in $T^{(r)}_i$ and $T^{(\infty)}_i$ when $r \geq \ell$. Consequently, our preceding proof remains valid in this context: both

[5] The proof follows identical steps sicne in the proof of Lemma 5 we used tightness of the graph structure without considering the marks corresponding to the epidemics. Here we can use the same arguments.

the variance bound on $\mathbb{1}\{t < T_1^{(r)}\}$ and the convergence to the limit applies to $D^{(i)}(t)$. As a result, $D_n^{(i)}(t) \xrightarrow{\mathbb{P}} D^{(i)}(t)$.

Then the conclusion follows by noting that the number of nodes that in states $\mathcal{D}_i$, can be obtained by the following subtraction $D_n^{(i)}(t) - D_n^{(i-1)}(t)$.

D.2. Proof of Corollary 2

Our goal is to prove the convergence of the epidemics with a generalized starting configuration. We show how Lemmas 1 and 4 can be extended to this case. As before, note that $S_n(t) - S_{n,r}(t)$ only includes nodes that $T^{(\infty)}(o_n) \neq T^{(r)}(o_n)$. For such nodes, there exists a path of length at least $r+1$ from v to an initially infected node.

$$\mathbb{E}_{\Xi}[\sup_{t\geq 0} |\mathscr{E}_n(t) - \mathscr{E}_{n,r}(t)|] = \mathbb{P}_{\Xi}\Big(T^{(\infty)}(o_n) \neq T^{(r)}(o_n) \mid \mathcal{M}(G_n)\Big),$$

where the randomness on the right-hand side is over uniform random node o_n and the infection marks on the initial graph. With the local reachability condition in hand, the right-hand side tends to 0 as we increase n and then r.

The second subtle difference in the proof of Theorem 4 is that in the variance bound in Lemma 3, we have used the fact that $T^{(r)}(u)$ and $T^{(r)}(v)$ are independent if u and v have distance larger than $2r$. With the generalized starting configuration, the independence still holds if u and v have a distance larger than $2r + \ell$. Recall that ℓ is the neighborhood size that initial conditions depend on (through the function P_ℓ). The rest of the proof follows as those of Theorems 3 and 4 as in other parts of the proof, we do not use the starting configuration.

Appendix E: Proof of Theorem 5 - Variance and Bias bounds of Weighted Estimator

Bounding the bias: First, we show the bias is the same as the estimator with uniform random queries. Fix $t \geq 0$. Let $v_1, \dots, v_q$ be the q i.i.d. draws from the weighted distribution $(p_i)_{i\in[n]}$. Taking expectation over both the random draws $\{v_1, \dots, v_q\}$ while conditioning on the epidemic process (i.e., the marks) shows

$$\mathbb{E}\Big[\widehat{S}_{q,k,n,\vec{p}}(t)\,\Big|\,\mathcal{M}(G_n)\Big] \;=\; \frac{1}{n\,q}\sum_{j=1}^{q}\mathbb{E}\Big[\tfrac{S_{k,v_j}(t)}{p_{v_j}}\,\Big|\,\mathcal{M}(G_n)\Big] \;=\; \frac{1}{n}\sum_{i=1}^{n} S_{k,i}(t),$$

since each v_i is chosen with probability p_i. But $\frac{1}{n}\sum_{i=1}^{n} S_{k,i}(t)$ is exactly the expected value of the uniform-sampling estimator that measures how many nodes are susceptible when each node is sampled with probability $1/n$. Hence the two estimators share the same expectation, completing the proof. Thus,

$$\mathbb{E}\Big[\widehat{S}_{q,k,n,\vec{p}}(t)\Big] \;=\; \mathbb{E}\Big[\hat{S}^{(\rho)}_{q,k,n}(t)\Big],$$

where the expectation is over the randomness of the choice of q queries. Now, if we take expectation over the randomness of the graph and epidemics as well, using the proof of Theorem 1 we get, for any $t \geq 0$, and any r, the following inequalities:

$$\mathbb{E}\Big[\hat{S}^{(\rho)}_{q,k,n}(t)\Big] - \mathbb{E}\Big[S^{(\rho)}_{n,r}(t)\Big] \leq \varepsilon_r(G_n, k),$$

and

$$\mathbb{E}\Big[\hat{S}^{(\rho)}_{q,k,n}(t)\Big] \geq \mathbb{E}\Big[S^{(\rho)}_{n}(t)\Big] \qquad \text{and} \qquad \mathbb{E}\Big[\hat{S}^{(\rho)}_{n,r}(t)\Big] \geq \mathbb{E}\Big[S^{(\rho)}_{n}(t)\Big].$$

Then, by applying Lemma 1, and the triangle inequality, for any choice of r,

$$\sup_{t\geq 0} |\mathbb{E}\Big[\widehat{S}_{q,k,n,\vec{p}}(t)\Big] - \mathbb{E}\Big[S_n(t)\Big]| \leq (1-\rho)^r + \varepsilon_r(G_n, k),$$

which finishes the proof of this part. □

Bounding the variance: We outline the key steps mirroring the proof techniques used in Lemmas 2 and 3 (which treat the uniform-sampling case).

First consider the case of a deterministic graph.

Step 1. Bound the conditional variance given the graph. Let $X_i(t) := \frac{S_{k,i}(t)}{np_i}$. If $\mathrm{dist}(u,v) \geq 2k$, then X_u and X_v are independent. Thus,

$$\begin{aligned}\mathrm{Var}\Big(\tfrac{1}{n}\sum_{j=1}^{n} X_{v_j}(t)\Big) &\leq \frac{1}{n^2}\sum_{j=1}^{q}\mathrm{Var}(X_{v_j}(t)) + \frac{1}{n^2}\sum_{u,v:\mathrm{dist}(u,v)\leq 2k} p_u p_v Cov(X_v(t), X_u(t)) \\ &\leq \frac{1}{n\min_{i\in[n]} np_i} + \Big(\frac{\max_{i\in[n]} p_i}{\min_{i\in[n]} p_i}\Big)^2 \varepsilon_{2k}(G_n).\end{aligned}$$

Now we need to bound it for q nodes. Conditioned on the marks (i.e., when only the choice of the initial nodes is random) the X_{v_i} are independent. Then, by the Azuma-Hoeffding bound,

$$Var(\frac{1}{q}\sum_{i=1}^{q} X_{v_i}(t)|\mathcal{M}(G_n)) \leq \frac{1}{q(\min_{i\in[n]} np_i)}.$$

We now use the fact that

$$\begin{aligned}\mathrm{Var}\big(\widehat{S}_{q,k,n,\vec{p}}(t)\big) &= \mathbb{E}\Big[\mathrm{Var}\Big(\widehat{S}_{q,k,n,\vec{p}}(t)\,\big|\,\mathcal{M}(G_n)\Big)\Big] + \mathrm{Var}\Big(\mathbb{E}\big[\widehat{S}_{q,k,n,\vec{p}}(t)\,\big|\,\mathcal{M}(G_n)\big]\Big). \\ &= \mathbb{E}\Big[\mathrm{Var}\Big(\tfrac{1}{q}\sum_{j=1}^{q} X_{v_j}(t)\,\big|\,\mathcal{M}(G_n)\Big)\Big] + \mathrm{Var}\Big(\tfrac{1}{n}\sum_{j=1}^{n} X_{v_j}(t)\Big) \\ &\leq \frac{1}{q\min_{i\in[n]} np_i} + \frac{1}{n\min_{i\in[n]} p_i} + \Big(\frac{\max_{i\in[n]} p_i}{\min_{i\in[n]} p_i}\Big)^2 \varepsilon_{2k}(G_n).\end{aligned}$$

Step 2. Bound on variance for the random graphs. We decompose the variance by conditioning on G_n to obtain

$$\mathrm{Var}\big(\widehat{S}_{q,k,n,\vec{p}}(t)\big) = \mathbb{E}\Big[\mathrm{Var}\Big(\widehat{S}_{q,k,n,\vec{p}}(t)\,\big|\,G_n\Big)\Big] + \mathrm{Var}\Big(\mathbb{E}\big[\widehat{S}_{q,k,n,\vec{p}}(t)\,\big|\,G_n\big]\Big).$$

The first term can be bounded by step 1. Next, we examine $\mathrm{Var}\big(\mathbb{E}[\widehat{S}_{q,k,n,\vec{p}}(t)\,|\,G_n]\big)$. Here,

$$\mathbb{E}\big[\widehat{S}_{q,k,n,\vec{p}}(t)\,\big|\,G_n\big] = \frac{1}{n}\sum_{i=1}^{n} S_{k,i}(t) = \mathbb{E}\big[\widehat{S}_{q,k,n}(t)\,\big|\,G_n\big]$$

where the right-hand side is the estimator with uniform random initial seeds. Thus, by applying Lemma 3, for any $\epsilon > 0$, there exists N_0 such that if $n > N_0$ then

$$\mathrm{Var}\Big(\mathbb{E}[\widehat{S}_{q,k,n,\vec{p}}(t)\,\big|\,G_n]\Big) \leq \epsilon.$$

By combining the two steps we get the results of the theorem. □

Appendix F: Proof of Theorem 6 – Robustness to Misspecified Distributions

Recall that $T_r(G_n, v, \mathcal{M}(G_n))$ denotes the infection time of a node v under the assumption that the epidemic process is confined to the ball $B_r(G_n, v)$. Our goal is to study the probability of the event $\{T_r(G_n, v, \mathcal{M}(G_n)) > t\}$ and bound the difference in this probability when the contact and recovery times are drawn from the true distributions (D_I, D_R) versus the approximate distributions $(\widetilde{D}_I, \widetilde{D}_R)$.

For this purpose, we first set the notation up. As before, let $c_{(u,w)}$ be the random contact time on edge (u,w) drawn from D_I, and let $\widetilde{c}_{(u,w)}$ be the corresponding contact time drawn from $\widetilde{D}_I$. Similarly, let r_u be the recovery time of node u from D_R, and let $\widetilde{r}_u$ be the recovery time from $\widetilde{D}_R$. Define

$$c'_{(u,w)} = \begin{cases} c_{(u,w)}, & \text{if } c_{(u,w)} < r_u, \\ \infty, & \text{otherwise,} \end{cases}$$

and similarly,

$$\tilde{c}'_{(u,w)} = \begin{cases} \widetilde{c}_{(u,w)}, & \text{if } \widetilde{c}_{(u,w)} < \widetilde{r}_u, \\ \infty, & \text{otherwise.} \end{cases}$$

Intuitively, $c'_{(u,w)}$ is the time at which u can infect w along edge (u,w), assuming u is not yet recovered; if u recovers before transmitting, we set $c'_{(u,w)} = \infty$.

Given a possible transmission path $(v_1, v_2, \ldots, v_k = v)$ within $B_r(G_n, v)$, the time that it takes to infect v through this path is

$$\sum_{i=1}^{k-1} c'_{(v_i,v_{i+1})} \quad \text{or} \quad \sum_{i=1}^{k-1} \tilde{c}'_{(v_i,v_{i+1})}$$

under the true or approximate system, respectively. Thus,

$$T_r(G_n, v, \mathcal{M}(G_n)) = \min_{\substack{\text{paths } p:\\ p\subseteq B_r(G_n,v)}} \sum_{(u,w)\in p} c'_{(u,w)},$$

and the analogous quantity $\widetilde{T}_r(G_n, v, \mathcal{M}(G_n))$ uses $\tilde{c}'$-variables.

Now, turning our attention to our goal of bounding

$$\Big|\mathbb{P}\big(T_r(G_n,v,\mathcal{M}(G_n))>t\big)\;-\;\mathbb{P}\big(\widetilde{T}_r(G_n,v,\mathcal{M}(G_n))>t\big)\Big|.$$

If $T_r(G_n,v,\mathcal{M}(G_n))>t$, it implies every path in $B_r(G_n,v)$ has $\sum_{(u,w)\in p}c'_{(u,w)}>t$. Equivalently, for $\widetilde{T}_r$, every path must have $\sum_{(u,w)\in p}\tilde{c}'_{(u,w)}>t$. As a result,

$$\begin{aligned}&\Big|\mathbb{P}\big(T_r(G_n,v,\mathcal{M}(G_n))>t\big)\;-\;\mathbb{P}\big(\widetilde{T}_r(G_n,v,\mathcal{M}(G_n))>t\big)\Big|\\&\quad=\Big|\mathbb{P}\big(\sum_{(u,w)\in p}c'_{(u,w)}\geq t\,\forall\text{ paths }p\subseteq B_r(G_n,v)\big)\;-\;\mathbb{P}\big(\sum_{(u,w)\in p}\tilde{c}'_{(u,w)}\geq t\,\forall\text{ paths }p\subseteq B_r(G_n,v)\big)\Big|.\end{aligned}$$

Now using the alternative functional definition of total variation distance and data processing inequality (see, e.g., (Beaudry and Renner 2012)), we can bound the difference in these events is at most the total variation distance over all random variables $\{c'_{(u,w)},r_u\}$ versus $\{\tilde{c}'_{(u,w)},\tilde{r}_u\}$, i.e.,

$$\Big|\mathbb{P}\big(T_r(G_n,v,\mathcal{M}(G_n))>t\big)\;-\;\mathbb{P}\big(\widetilde{T}_r(G_n,v,\mathcal{M}(G_n))>t\big)\Big|\leq d_{\mathrm{TV}}(\{\tilde{c}'_{(u,w)},\tilde{r}_u\forall u,v\},\{c'_{(u,w)},r_u\forall u,v\}).$$

Here the nodes and edges are chosen within the ball $B_r(G_n,v)$. Since c' is a function of c and r, again by applying the data-processing inequality,

$$d_{\mathrm{TV}}(\{\tilde{c}'_{(u,w)},\tilde{r}_u\forall u,v\},\{c'_{(u,w)},r_u\forall u,v\})\leq d_{\mathrm{TV}}(\{\tilde{c}_{(u,w)},\tilde{r}_u\forall u,v\},\{c_{(u,w)},r_u\forall u,v\}).$$

The total variation bound across all contact and recovery times in the ball can be crudely bounded by summing pairwise TV distances:

$$d_{\mathrm{TV}}(\{\tilde{c}_{(u,w)},\tilde{r}_u\forall u,v\},\{c_{(u,w)},r_u\forall u,v\})\leq\sum_{(u,w)\in B_r(G_n,v)}d_{\mathrm{TV}}\big(c_{(u,w)},\widetilde{c}_{(u,w)}\big)+\sum_{u\in B_r(G_n,v)}d_{\mathrm{TV}}\big(r_u,\widetilde{r}_u\big).$$

Since $d_{\mathrm{TV}}\big(c_{(u,w)},\widetilde{c}_{(u,w)}\big)\leq\epsilon_I$ and $d_{\mathrm{TV}}\big(r_u,\widetilde{r}_u\big)\leq\epsilon_R$ by assumption, we obtain

$$\leq|E(B_r(G_n,v))|\cdot\epsilon_I+|V(B_r(G_n,v))|\cdot\epsilon_R\leq(|B_r(G_n,v)|+|B_r(G_n,v)|^2)(\epsilon_I+\epsilon_R).$$

Now, if we sum this up for all nodes, then for uniform random queries the error for estimating infection time is bound by:

$$\begin{aligned}d_{\mathrm{TV}}\Big(\tilde{\mathscr{E}}^{(\rho)}_{n,q,k}(t),\hat{\mathscr{E}}^{(\rho)}_{n,q,k}(t)\Big)&\leq\frac{1}{n}\Big(\sum_{v\in V(G_n)}\mathbb{1}\{|B_r(G_n,v)|<k\}(k+k^2)(\epsilon_I+\epsilon_R)+\mathbb{1}\{|B_r(G_n,v)|\geq k\}\Big)\\&\leq\big((k+k^2)(\epsilon_I+\epsilon_R)\big)(1-\varepsilon_r(G_n,k))+\varepsilon_r(G_n,k).\end{aligned}$$

The same bound would hold for averaging over q queries, hence

$$d_{\mathrm{TV}}\Big(\tilde{\mathscr{E}}^{(\rho)}_{n,q,k}(t),\hat{\mathscr{E}}^{(\rho)}_{n,q,k}(t)\Big)\leq\big((k+k^2)(\epsilon_I+\epsilon_R)\big)(1-\varepsilon_r(G_n,k))+\varepsilon_r(G_n,k).$$

Appendix G: Additional Results on Empirical Validation

Details on Synthetic Network Creation The Preferential Attachment model represents the growth of scale-free networks, where new nodes join the network and preferentially connect to existing nodes based on their degrees. In our experiments, we set the parameter $m = 3$, indicating that each new node forms exactly three connections with existing nodes.

For Random Geometric Graphs, nodes are randomly distributed in a Euclidean space with their positions defined by x and y axes drawn uniformly at random from $[0, \sqrt{n}]$, where n denotes the size of the graph. The connection radius for the Random Geometric Graph is set to 1.5, ensuring an average degree of approximately 7.06 as $n \to \infty$.

Details on Evaluating Estimator's Error: To assess the performance of our estimator, we compare its output against the ground-truth SIR process (ran on the whole network, and averaged for 1000 simulations) using the following error metrics (for $q = 10$):

- *Absolute Error of the Final Epidemic Size:* $|R_n(\infty) - \hat{R}_{q,k,n}(\infty)|$, which quantifies the deviation of the estimated final epidemic size from the ground truth (values in $[0,1]$).
- *Euclidean Distance of the Time Evolution:*

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \|\mathscr{E}_n(t) - \hat{\mathscr{E}}_{q,k,n}(t)\|_1 \, dt,$$

measuring the average deviation between the estimated and actual epidemic trajectories.

- *Pearson Correlation of the Time Evolution:* This statistic measures the linear correlation between the estimated and ground-truth time series; values closer to 1 indicate a stronger match.

Tables 1, 2, 3, and 4 report these metrics (with 95% confidence intervals) for various values of k (the probing budget per query). Additionally, Table 5 illustrates the effect of increasing graph size on the Euclidean distance error for both Preferential Attachment and Random Geometric Graphs. Note the incremental benefit of increasing k diminishes. In practice, one can select k via pilot experiments by monitoring how the estimated time evolution changes with k; once further increases yield only negligible improvements relative to the target accuracy, k can be fixed at that level.

Misspecified Distribution: Let $\{\epsilon_i\}_{i=1}^{N=1000}$ denote the errors between the ith trial of the estimator for the final epidemic size, $\hat{R}^{(i)}_{q,k,n}(\infty)$, and the ground-truth final epidemic size, $\mathbb{E}\big[R(\infty)\big]$, i.e.,

$$\epsilon_i = \hat{R}^{(i)}_{q,k,n}(\infty) - \mathbb{E}\big[R(\infty)\big].$$

Then, the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are defined as

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} \left|\hat{R}_i^{(\infty)} - R^{(\infty)}\right|, \qquad \text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{R}_i^{(\infty)} - R^{(\infty)}\right)^2}. \tag{25}$$

Table 6 reports these error metrics along with the associated 95% confidence intervals for three different scenarios: (i) using the true distributions, where D_I is an exponential random variable with rate 1; (ii) using approximations $\widetilde{D}_I$ obtained by scaling the true parameters by 0.9, which corresponds to a total variation distance of $d_{\mathrm{TV}}(\widetilde{D}_I, D_I) = 0.039$; and (iii) using approximations $\widetilde{D}_I$ obtained by scaling the true parameters by 0.8, which corresponds to a total variation distance of $d_{\mathrm{TV}}(\widetilde{D}_I, D_I) = 0.082$. In all cases, the estimated final epidemic size is compared against the ground truth, $\mathbb{E}\big[R(\infty)\big]$, which is computed using the true distributions D_I and D_R. This sensitivity analysis serves as a test for Theorem 6.

Weighted Queries Sampling and the HT Estimator: In this additional experiment, we test the performance of the weighted Horvitz–Thompson (HT) estimator (see Section 2.3, and Theorem 5). Instead of selecting the $q = 10$ query nodes uniformly at random, we choose them with probabilities proportional to their degree plus 0.01 (to ensure that isolated nodes are included). For each trial, we compute the error

$$\epsilon_i = \hat{R}^{(i)}_{\vec{p},q,k,n}(\infty) - \mathbb{E}[R(\infty)],$$

where $\hat{R}^{(i)}_{\vec{p},q,k,n}(\infty)$ is the HT estimator of the final epidemic size from the ith trial and $\mathbb{E}[R(\infty)]$ is the ground-truth final epidemic size.

Figure 8 displays the histogram of the ϵ_i values for the Copenhagen dataset under weighted sampling and uniform sampling, respectively. For the San Francisco dataset, analogous experiments were done see Figure 9. In these plots, rows correspond to different initial infection probabilities ρ and columns correspond to different probing budgets k (with $q = 10$ queries). While the uniform sampling yields slightly smaller RMSE and MAE values for the Copenhagen dataset, the 95% confidence interval bounds remain tight in both cases, demonstrating the robustness of our estimator under either sampling strategy.

Vaccination Experiments: To assess the performance of our estimator under more general epidemic settings, we conducted experiments in which a fixed fraction of nodes is vaccinated at time 0. Specifically, each node is vaccinated with probability ρ_V; non-vaccinated nodes follow the standard initialization (i.e., they are susceptible with probability $1 - \rho$ and infected otherwise). The epidemic then evolves according to the multi-stage time-varying model, with nodes transitioning through infection stages independently (see Example 1).

We apply the estimator as in our standard experiments to predict both the time evolution and the final epidemic size. Table 7 summarizes key performance metrics—using the previously defined measures—for different vaccination fractions. The results indicate that our estimator remains robust in the presence of vaccination.

Table 1 Performance evaluation of the estimator on Copenhagen Dataset. Values in parentheses represent confidence intervals (CI).

Probing Budget k	Final Size	Time Evolution	Time Evolution
	Absolute Error (CI)	Euclidean Distance (CI)	Pearson Correlation of (CI)
2	0.091 (0.078, 0.104)	0.011 (0.010, 0.012)	0.942 (0.932, 0.952)
3	0.084 (0.071, 0.097)	0.010 (0.010, 0.011)	0.945 (0.937, 0.953)
4	0.086 (0.073, 0.100)	0.010 (0.010, 0.011)	0.952 (0.944, 0.959)
5	0.095 (0.079, 0.111)	0.010 (0.009, 0.010)	0.955 (0.947, 0.962)
6	0.089 (0.074, 0.104)	0.010 (0.010, 0.011)	0.959 (0.953, 0.965)
7	0.106 (0.091, 0.121)	0.010 (0.010, 0.011)	0.959 (0.953, 0.966)
8	0.087 (0.074, 0.099)	0.010 (0.009, 0.011)	0.964 (0.959, 0.969)
9	0.084 (0.071, 0.097)	0.010 (0.009, 0.010)	0.965 (0.959, 0.970)

Table 2 Performance evaluation of the estimator on San Francisco Dataset. Values in parentheses represent confidence intervals (CI).

Probing Budget k	Final Size	Time Evolution	Time Evolution
	Absolute Error (CI)	Euclidean Distance (CI)	Pearson Correlation of (CI)
2	0.103 (0.086, 0.120)	0.058 (0.053, 0.063)	0.595 (0.531, 0.659)
3	0.091 (0.078, 0.105)	0.056 (0.052, 0.060)	0.560 (0.494, 0.625)
4	0.117 (0.099, 0.134)	0.058 (0.053, 0.063)	0.652 (0.598, 0.707)
5	0.107 (0.092, 0.122)	0.060 (0.056, 0.064)	0.542 (0.476, 0.607)
6	0.114 (0.100, 0.127)	0.059 (0.054, 0.065)	0.538 (0.467, 0.610)
7	0.115 (0.098, 0.131)	0.061 (0.056, 0.065)	0.554 (0.494, 0.615)
8	0.105 (0.091, 0.118)	0.062 (0.056, 0.067)	0.510 (0.437, 0.582)
9	0.098 (0.085, 0.111)	0.057 (0.054, 0.061)	0.582 (0.509, 0.654)

Table 3 Performance evaluation of the estimator on Preferential Attachment with 500 nodes. Values in parentheses represent confidence intervals (CI).

Probing Budget k	Final Size Absolute Error (CI)	Time Evolution Euclidean Distance (CI)	Time Evolution Pearson Correlation of (CI)
2	0.070 (0.059,0.079)	0.013 (0.013, 0.014)	0.690 (0.657, 0.723)
3	0.063 (0.050,0.074)	0.012 (0.011, 0.013)	0.718 (0.679, 0.757)
4	0.079 (0.065,0.092)	0.012 (0.011, 0.013)	0.718 (0.680, 0.757)
5	0.079 (0.066,0.091)	0.011 (0.011, 0.012)	0.759 (0.720, 0.799)
6	0.073 (0.062,0.083)	0.011 (0.010, 0.012)	0.758 (0.718, 0.797)
7	0.072 (0.061,0.081)	0.011 (0.010, 0.011)	0.776 (0.742, 0.810)
8	0.066 (0.056,0.075)	0.011 (0.011, 0.012)	0.759 (0.729, 0.789)
9	0.067 (0.057,0.074)	0.010 (0.010, 0.011)	0.791 (0.763, 0.818)

Table 4 Performance evaluation of the estimator on Random Geometric Graph with 500 nodes. Values in parentheses represent confidence intervals (CI).

Probing Budget k	Final Size Absolute Error (CI)	Time Evolution Euclidean Distance (CI)	Time Evolution Pearson Correlation of (CI)
2	0.080 (0.070, 0.089)	0.063 (0.052, 0.073)	0.855 (0.835, 0.876)
3	0.100 (0.084, 0.11)	0.090 (0.078, 0.102)	0.836 (0.809, 0.864)
4	0.075 (0.063, 0.085)	0.068 (0.058, 0.077)	0.851 (0.830, 0.871)
5	0.075 (0.064, 0.085)	0.077 (0.066, 0.088)	0.864 (0.844, 0.884)
6	0.074 (0.063, 0.084)	0.083 (0.071, 0.095)	0.839 (0.812, 0.865)
7	0.089 (0.078, 0.100)	0.066 (0.056, 0.076)	0.844 (0.820, 0.868)
8	0.076 (0.063, 0.082)	0.074 (0.063, 0.086)	0.863 (0.845, 0.881)
9	0.084 (0.074, 0.092)	0.071 (0.059, 0.084)	0.878 (0.862, 0.894)

Table 5 Absolute error of estimator of the final size of the epidemic the for growing graph size (n). The number of queries is 10, and the testing budget per query is $k = 4$. Confidence intervals are obtained with 1000 simulations.

Graph Size (n)	Preferential Attachment	Random Geometric Graph
	Euclidean Distance (CI)	Euclidean Distance (CI)
500	0.079 (0.066, 0.091)	0.075 (0.064, 0.085)
1000	0.076 (0.064, 0.088)	0.072 (0.061, 0.083)
2000	0.076 (0.063, 0.090)	0.071 (0.061, 0.082)
5000	0.074 (0.064, 0.085)	0.067 (0.059, 0.075)
10000	0.068 (0.058, 0.077)	0.066 (0.058, 0.074)

Table 6 Sensitivity Analysis with Respect to Distribution Misspecification. The table tests the robustness result of Theorem 6 using the true distributions (D_I, exponential with rate 1), and approximations $\widetilde{D}_I$ obtained by scaling the true parameters by 0.9 and 0.8. We fixed $\rho = 0.01$, $k = 5$ and $q = 10$.

Copenhagen Data			
Estimator	**MAE**	**RMSE**	**95% CI**
True: $\tilde{R}_{q,k,n}(\infty)$ (D_I, D_R)	0.094	0.125	(-0.035, 0.015)
Slightly Misspecified: $\tilde{R}_{q,k,n}(\infty)$ $(0.9\times\ D_I, D_R)$	0.102	0.133	(-0.045, -0.025)
Moderately Misspecified: $\tilde{R}_{q,k,n}(\infty)$ $(0.8\times\ D_I, D_R)$	0.108	0.140	(-0.046, -0.029)
San Francisco Data			
Estimator	**MAE**	**RMSE**	**95% CI**
True: $\hat{R}_{q,k,n}(\infty)$ (D_I, D_R)	0.103	0.124	(-0.011, 0.021)
Slightly Misspecified: $\tilde{R}_{q,k,n}(\infty)$ $(0.9\times\ D_I, D_R)$	0.114	0.138	(-0.017, 0.004)
Moderately Misspecified: $\tilde{R}_{q,k,n}(\infty)$ $(0.8\times\ D_I, D_R)$	0.121	0.144	(-0.038, 0.014)

Table 7 Performance of the estimator with vaccination for Copenhagen and San Francisco datasets. Values in parentheses represent 95% confidence intervals. We fixed $\rho = 0.01$, $k = 5$ and $q = 10$.

Copenhagen Data

Vaccination	Final Size	Time Evolution	Time Evolution
Fraction ρ_V	Abs. Error (CI)	Euclidean Distance (CI)	Pearson Correlation (CI)
0.05	0.085 (0.070, 0.100)	0.012 (0.011, 0.013)	0.950 (0.940, 0.960)
0.10	0.092 (0.078, 0.106)	0.013 (0.012, 0.014)	0.945 (0.935, 0.955)
0.20	0.101 (0.088, 0.114)	0.015 (0.014, 0.016)	0.940 (0.930, 0.950)

San Francisco Data

Vaccination	Final Size	Time Evolution	Time Evolution
Fraction ρ_V	Abs. Error (CI)	Euclidean Distance (CI)	Pearson Correlation (CI)
0.05	0.095 (0.080, 0.110)	0.058 (0.053, 0.063)	0.715 (0.671, 0.789)
0.10	0.088 (0.075, 0.101)	0.056 (0.052, 0.060)	0.672 (0.589, 0.742)
0.20	0.105 (0.092, 0.118)	0.060 (0.056, 0.064)	0.676 (0.592, 0.745)

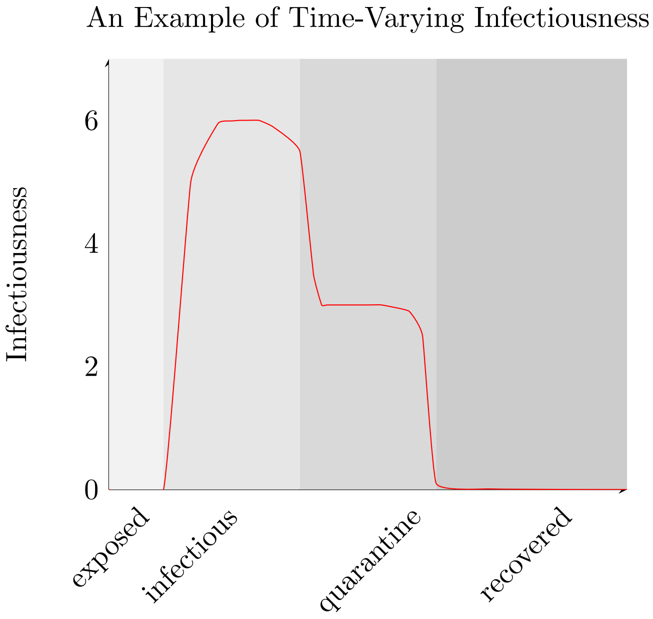

Figure 1 The density of infectiousness over time, illustrating the intervals of exposure, infectiousness, quarantine with reduced transmission rate, and recovery.

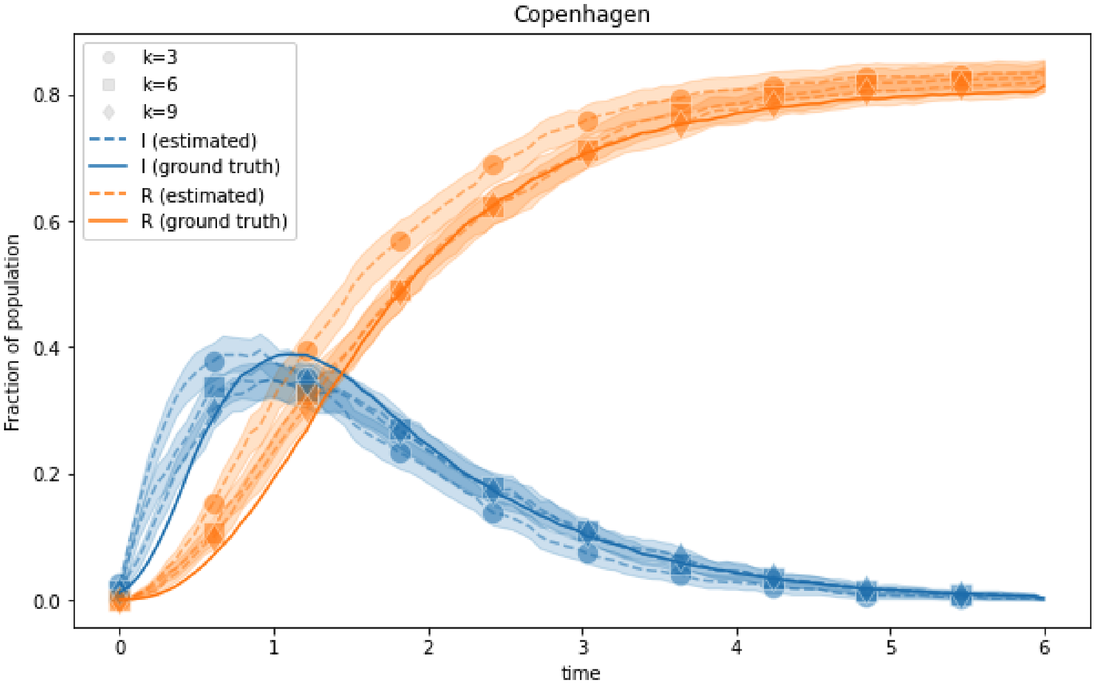

Figure 2 Copenhagen Dataset: Time Evolution of Epidemic Infections (I) and Recoveries (R) with Estimator Evolution for Various Testing Budgets (k)

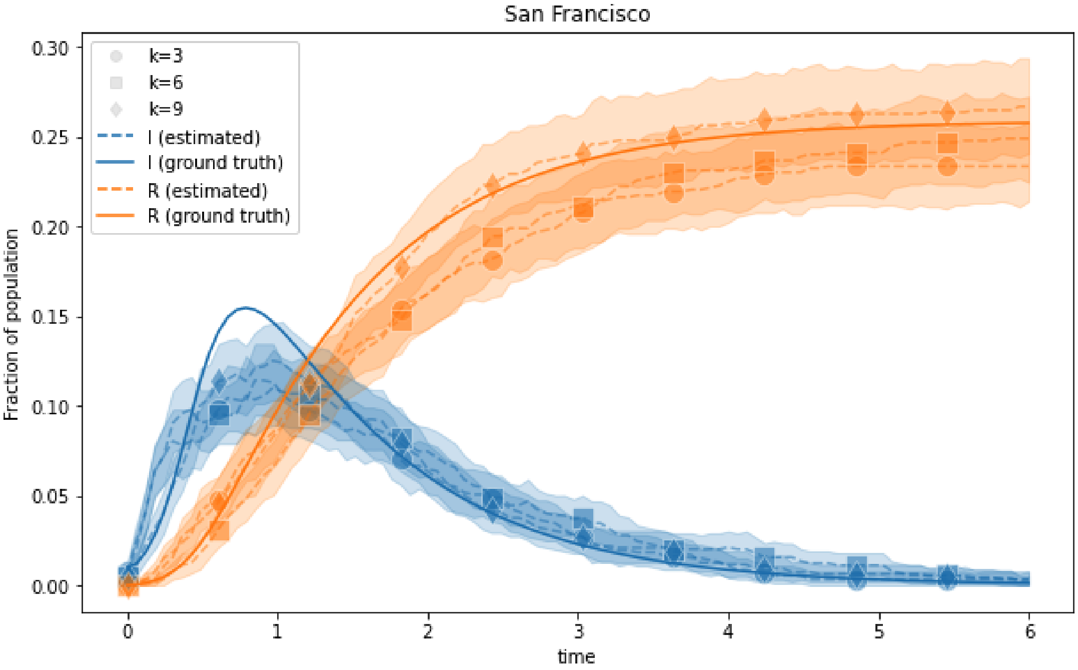

Figure 3 San Francisco Dataset: Time Evolution of Epidemic Infections (I) and Recoveries (R) with Estimator Evolution for Various Testing Budgets (k)

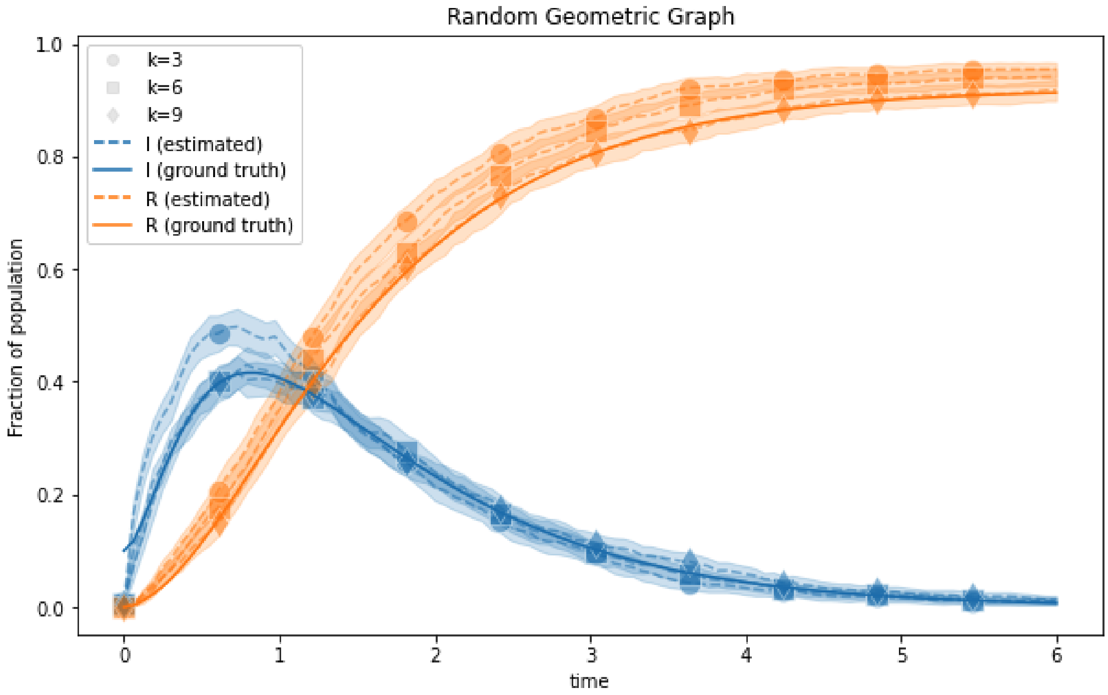

Figure 4 Random Geometric Graph: Time Evolution of Epidemic Infections (I) and Recoveries (R) with Estimator Evolution for Various Testing Budgets (k)

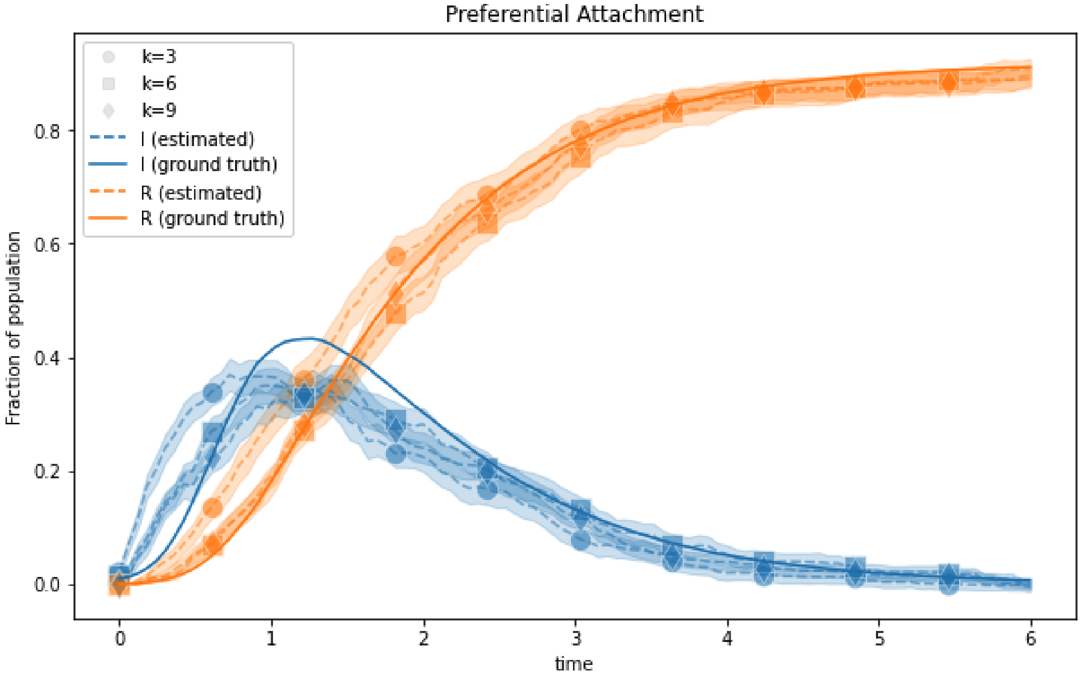

Figure 5 Preferential Attachment: Time Evolution of Epidemic Infections (I) and Recoveries (R) with Estimator Evolution for Various Testing Budgets (k)

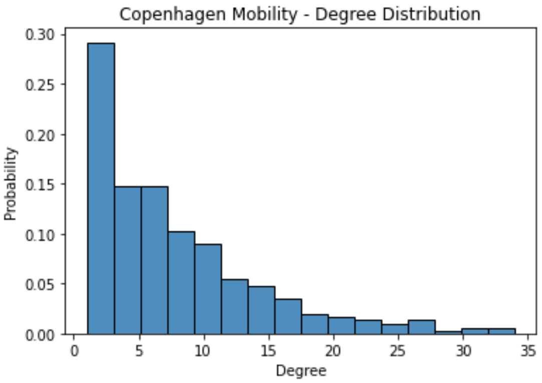

Figure 6 Degree Distribution of Copenhagen Network.

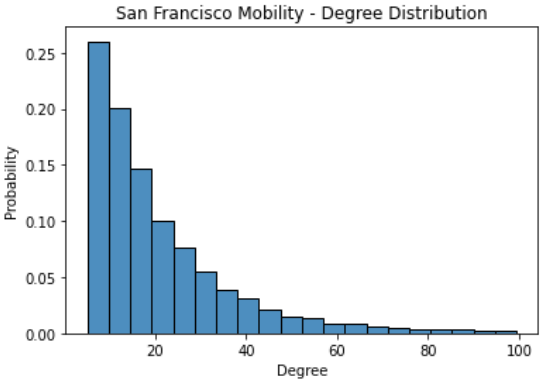

Figure 7 Degree Distribution of San Francisco Mobility Network (Capped at 100 for Depiction).

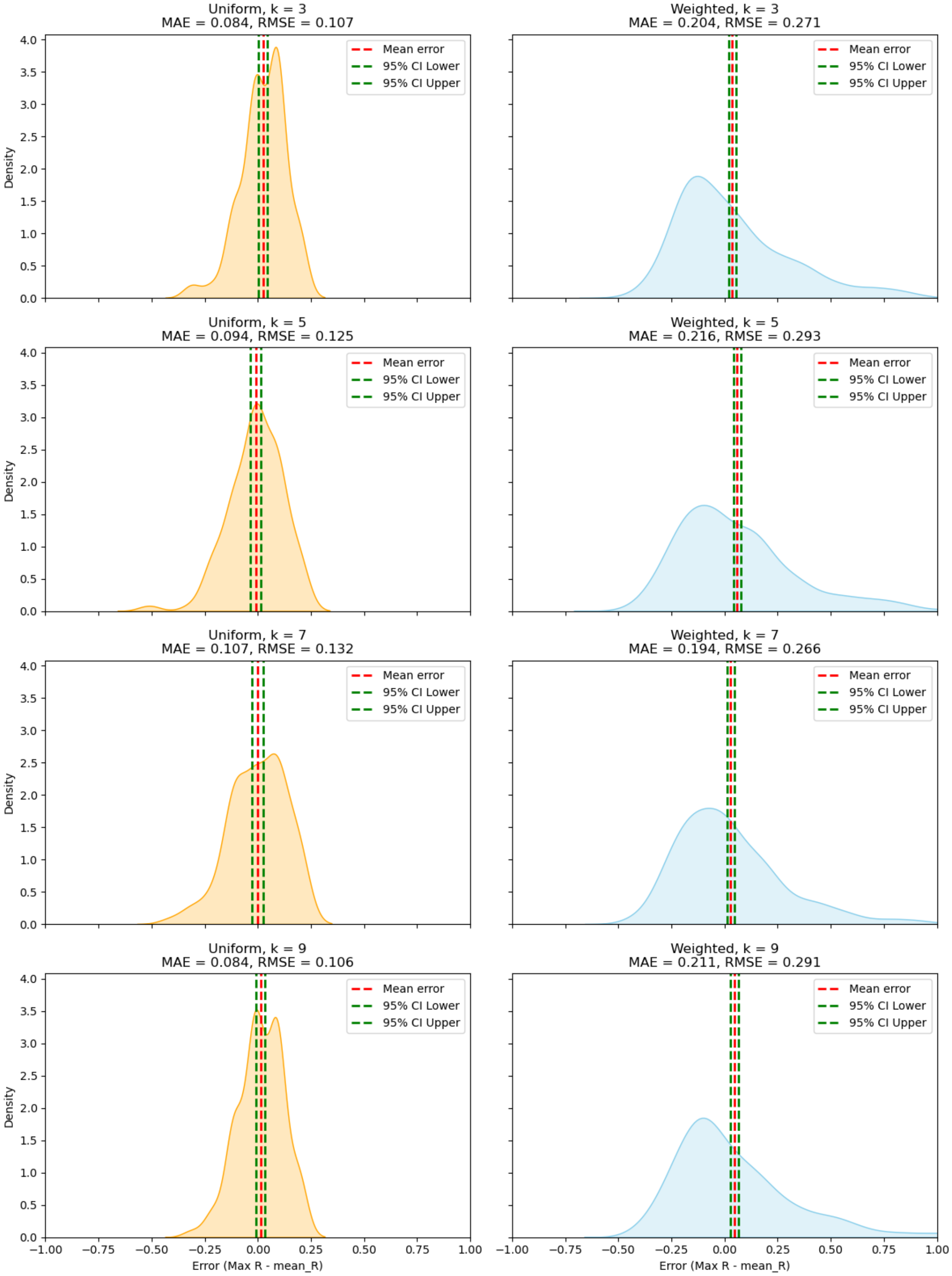

Figure 8 Error distributions for the final epidemic size $R(\infty)$ in **Copenhagen dataset** using two sampling strategies: uniform sampling (first column) and weighted sampling (second column, with selection probabilities proportional to node degree plus 0.01). Each plot shows the error distribution along with the mean error (red dashed line), the 95% confidence interval (green dashed lines), RMSE, and MAE. Rows correspond to different budgets k (with $q = 10$ queries per budget).

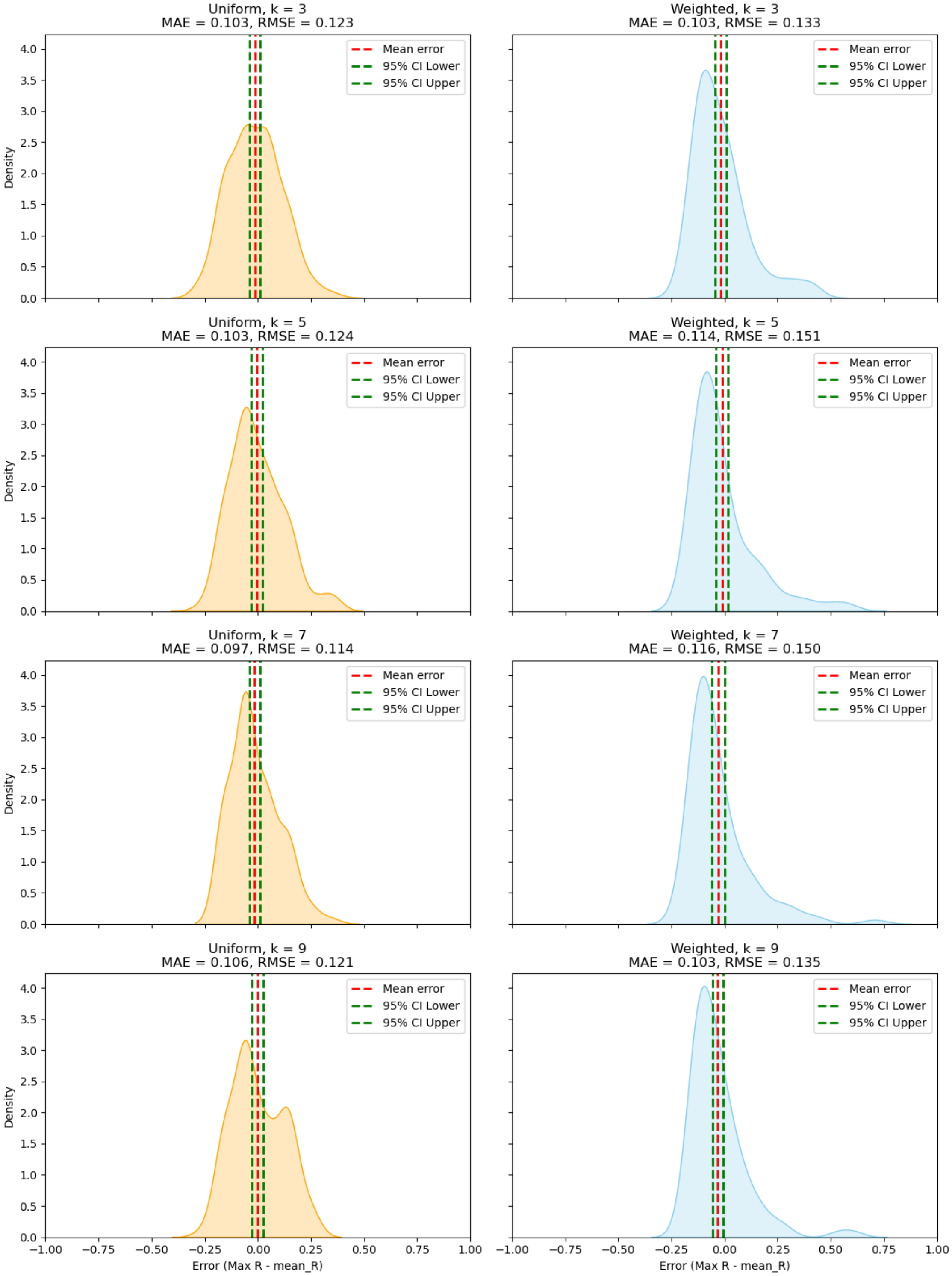

Figure 9 Error distributions for the final epidemic size $R(\infty)$ in **San Francisco dataset** using two sampling strategies: uniform sampling (first column) and weighted sampling (second column, with selection probabilities proportional to node degree plus 0.01). Each plot shows the error distribution along with the mean error (red dashed line), the 95% confidence interval (green dashed lines), RMSE, and MAE. Rows correspond to different budgets k (with $q = 10$ queries per budget).