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.,

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

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 $\mathrm{dist}_{G_n}(x,y) > 2r$. Let $Z_v(t) = \mathbf{1}_{\{t<T^{(r)}(G_n,v,\mathcal{M}(G_n))\}}$ and $\bar{Z}_v(t) = \mathbf{1}_{\{t<T^{(r)}(G_n,v,\mathcal{M}(G_n))\}} - \mathbb{E}[\mathbf{1}_{\{t<T^{(r)}(G_n,v,\mathcal{M}(G_n))\}}]$. Then, we can write

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

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

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

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 $P_r^{(G_n)}(H^\star) = \frac{1}{|V(G_n)|}\sum_{v\in V(G_n)} \mathbf{1}_{\{B_r(G_n,v)\simeq H^\star\}}$ 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

$$\begin{aligned} &\mathbb{E}_\Xi(S^{(\rho)}_{n,r}(t)|G_n) \\ &= \sum_{H^\star\in\mathcal{H}} P_r^{(G_n)}(H^\star)\mathbb{P}_\Xi(t < T^{(r)}(H^\star)|H^\star). \end{aligned} \tag{13}$$

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

$$\begin{aligned} &\mathbb{E}(S^{(\rho)}_{n,r}(t)|G_n) \leq \varepsilon_r(G_n, k) \\ &+ \sum_{H^\star\in\mathcal{H}_k} P_r^{(G_n)}(H^\star)\mathbb{P}_\Xi(t < T^{(r)}(H^\star)|H^\star). \end{aligned} \tag{14}$$

Then using tightness, given any $\delta > 0$ for large enough k and n, $\mathbb{P}(\varepsilon_r(G_n,k) \geq \delta) \leq 1-\delta$. Therefore, we can assume that $\mathrm{Var}(\varepsilon_r(G_n,k)) \leq 2\delta^2$. Also, using the fact that $S_{n,r}(t) \leq 1$, we get $\mathrm{Cov}(\varepsilon_r(G_n,k), \mathbb{E}(S^{(\rho)}_{n,r}(t)|G_n)) \leq 2\delta$. Let $\Delta_{H^\star} = P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star)$. Then we have,

$$\mathrm{Var}(\mathbb{E}(S^{(\rho)}_{n,r}(t)|G_n)) \leq 2\delta^2 + 2\delta + \mathbb{E}\Big[\Big(\sum_{H^\star:|H^\star|\leq k} \mathbb{P}_\Xi(t < T^{(r)}(H^\star)|H^\star)\Delta_{H^\star}\Big)^2\Big].$$

The expectation on the right hand side can be rewritten as which can be rewritten as

$$\begin{aligned}
\mathbb{E}\Big[\Big(\sum_{H^\star:|H^\star|\leq k} \mathbb{P}_\Xi(t < T^{(r)}(H^\star)|H^\star)\Delta_{H^\star}\Big)^2\Big] &= \mathbb{E}\Big[\sum_{H^\star:|H^\star|\leq k} \mathbb{P}_\Xi(t < T^{(r)}(H^\star)|H^\star)\Delta^2_{H^\star}\Big] \\
+2\sum_{H^\star:|H^\star|\leq k}\sum_{H^{\star\prime}:|H^{\star\prime}|\leq k} &\Big(\mathbb{P}_\Xi(t < T^{(r)}(H^\star)|H^\star)\mathbb{P}_\Xi(t < T^{(r)}(H^{\star\prime})|H^{\star\prime})\mathbb{E}\Big[\Delta_{H^\star}\Delta_{H^{\star\prime}}\Big]\Big) \\
&\leq \mathbb{E}\Big[\sum_{H^\star}\Delta^2_{H^\star}\Big] + 2\mathbb{E}\Big[\sum_{H^\star:|H^\star|\leq k}\sum_{H^{\star\prime}:|H^{\star\prime}|\leq k}|\Delta_{H^\star}\Delta_{H^{\star\prime}}|\Big].
\end{aligned}$$

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 $\delta' \leq \frac{\delta}{N_{r,k}}$, there exists N such that for all $n > N$, and for all $H^\star$, $\mathbb{E}[(P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star))^2] \leq \delta'$. Therefore,

$$\mathrm{Var}\Big(\mathbb{E}(S^{(\rho)}_{n,r}(t)|G_n)\Big) \leq 3\delta + 4\delta^2. \tag{15}$$

Putting this together with Lemma 2, we get

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

Finally, the result follows by applying (17) 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 $t \in [0,\infty]$,

$$\mathbb{P}_\Xi\Big(|S^{(\rho)}_{n,r}(t) - S^{(\rho)}_n(t)| \geq \delta'\Big) \leq \frac{(1-\rho)^r}{\delta'}. \tag{16}$$

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

Combining the previous three inequalities, we get that for any $t\in[0,\infty]$, $\mathbb{P}_\Xi\Big(|S_n^{(\rho)}(t)-\mathbb{E}_\Xi[S_n^{(\rho)}(t)]|\geq 2\delta'+(1-\rho)^r\Big)\leq \frac{1}{(\delta')^2}\Big(\frac{1}{n}+\varepsilon_{2r}(G_n)\Big)+\frac{(1-\rho)^r}{\delta'}$.

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

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

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,

$$\varepsilon_r(G_n)\leq \varepsilon_r(G_n,k)+\frac{k}{n}. \tag{17}$$

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

$$\begin{aligned} &\mathbb{P}_\Xi\Big(|S_{n,r}^{(\rho)}(t)-\hat{S}_{q,r,n}^{(\rho)}(t)|\geq\delta\Big)\leq \\ &\quad 2e^{-2q\delta^2}+\frac{16}{\delta^2}\Big(\frac{1}{n}+\varepsilon_{2r}(G_n)\Big) \end{aligned}$$

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

Now, let $u_1,u_2,\ldots,u_q$ be the set of initial nodes sampled independently by the algorithm. Then $S_{n,r}^{(\rho)}(t)=\frac{1}{n}\sum_{v\in V(G_n)}Z_v(t)$, and $\hat{S}_{q,r,n}^{(\rho)}(t)=\frac{1}{q}\sum_{i=1}^q Z_{u_i}(t)$. 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 $t \in [0, \infty]$

$$\mathbb{P}\Big(|S^{(\rho)}_{n,r}(t) - \hat{S}^{(\rho)}_{q,r,n}(t)| \geq \delta \mid \mathcal{M}(G_n)\Big) \leq 2e^{-2q\delta^2}, \tag{18}$$

where the probability is only over the randomness of the starting points of the algorithm u_i. Then, to conclude, 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 $S^{(\rho)}_{n,r}$ and the estimator $\hat{S}^{(\rho)}_{q,r,n}$ to show the corresponding marks. Our goal is to prove the following lower bound,

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

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

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

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

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

Now, we can apply (18) to bound the coupling between the algorithm and the epidemic, given that the marks are the same.

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

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

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

which proves our claim. Combining this with (17), we get,

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

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

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

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 (13), and that P_Ξ is the probability over the graph marks. Also, recall that $s_r^{(\rho)}(t) = \mu_\Xi\Big(\mathbf{1}_{\{t<T^{(r)}(G,o)\}}\Big)$. 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 $p_r(H^\star) = \mu(\mathbf{1}_{\{B_r(G,o)\simeq H^\star\}})$ 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

$$\sup_{t\geq 0}\big|\mathbb{E}_\Xi[S_{n,r}^{(\rho)}(t)] - s_r^{(\rho)}(t)\big| = \sup_{t\geq 0} \tag{19}$$

$$\left|\sum_{H^\star\in\mathcal{H}} P_\Xi\Big(t < T^{(r)}(H^\star)\Big)\Big(p_r^{(G_n)}(H^\star) - p_r(H^\star)\Big)\right|. \tag{20}$$

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

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

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 Lemma 1-3 we get that for any $\delta > 0$, and any r and n large enough, and any $t \in [0, \infty]$,

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

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 $r \geq 1$,

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

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 $n > N$,

$$\mathbb{P}\left(|\hat{\mathscr{E}}^{(\rho)}_{n,q_\delta,k_\delta}(t) - \mathscr{E}^{(\rho)}_n(t)| > \delta\right) \le \delta. \tag{21}$$

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 $t \in [0, \infty]$,

$$\lim_{r\to\infty} \lim_{n\to\infty} \mathbb{P}(\left|S^{(\rho)}_n(t) - S^{(\rho)}_{n,r}(t)\right| \ge \delta) = 0. \tag{22}$$

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 $o_n \in V(G_n)$,

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

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 $T^{(\infty)}(v) \ne T^{(r)}(v)$, 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 $T^{(\infty)}(v) \ne T^{(r)}(v)$ is bounded by $(1-\rho)^r$, and, hence, if we take the probability over the marks, we prove our claim. As a result,

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

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 (22).

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, $\big|\mathbb{E}_\Xi[S^{(\rho)}_{n,r}(t)] - s_r(t)\big| \xrightarrow{\mathbb{P}} 0$, 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 the variance bound on $\mathbf{1}_{\{t<T^{(r)}_1\}}$ and the convergence to the limit applies to $D^{(i)}(t)$. As a result, $D^{(i)}_n(t) \xrightarrow{\mathbb{P}} D^{(i)}(t)$.

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

[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.

D.2. Proof of Corollary 2

Our goal is to prove the convergence of the epidemics with a generalized starting configuration. We show how Lemma 1,4 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 3,4 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 $v_1, \ldots, v_q$ be the q i.i.d. draws from the weighted distribution $(p_i)_{i\in[n]}$. Taking expectation over both the random draws $\{v_1, \ldots, v_q\}$ while conditioning on the epidemic process (i.e., the marks) shows

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

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:

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

and $\mathbb{E}\left[\hat{S}^{(\rho)}_{q,k,n}(t)\right] \geq \mathbb{E}\left[S^{(\rho)}_{n}(t)\right]$ and $\mathbb{E}\left[\hat{S}^{(\rho)}_{n,r}(t)\right] \geq \mathbb{E}\left[S^{(\rho)}_{n}(t)\right]$. Then, by applying Lemma 1, and the triangle inequality, for any choice of r, $\sup_{t\geq 0} |\mathbb{E}\left[\widehat{S}_{q,k,n,\vec{p}}(t)\right] - \mathbb{E}\left[S_n(t)\right]| \leq (1-\rho)^r + \varepsilon_r(G_n, k)$, 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 $\text{dist}(u, v) \geq 2k$, then X_u and X_v are independent. Thus,

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

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

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

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

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

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,

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

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

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

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

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,

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

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.,

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

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,

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

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

$$\begin{aligned} &d_{\mathrm{TV}}(\{\tilde{c}_{(u,w)}, \tilde{r}_u \forall u, v\}, \{c_{(u,w)}, r_u \forall u, v\}) \\ &\quad \le \sum_{(u,w)\in B_r(G_n,v)} d_{\mathrm{TV}}\big(c_{(u,w)}, \widetilde{c}_{(u,w)}\big) \\ &\qquad + \sum_{u\in B_r(G_n,v)} d_{\mathrm{TV}}\big(r_u, \widetilde{r}_u\big). \end{aligned}$$

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

$$\begin{aligned} &\le |E(B_r(G_n, v))| \cdot \epsilon_I + |V(B_r(G_n, v))| \cdot \epsilon_R \\ &\le (|B_r(G_n, v)| + |B_r(G_n, v)|^2)(\epsilon_I + \epsilon_R). \end{aligned}$$

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

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

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

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

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

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{R}_i^{(\infty)} - R^{(\infty)}\right|,$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{R}_i^{(\infty)} - R^{(\infty)}\right)^2}. \tag{23}$$

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 $d_{\text{TV}}(\widetilde{D}_I, D_I) = 0.039$; and (iii) using approximations $\widetilde{D}_I$ obtained by scaling the true parameters by 0.8, which corresponds to a total variation distance of $d_{\text{TV}}(\widetilde{D}_I, D_I) = 0.082$. 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 both figures, the number of queries is fixed at ($q = 10$), and the initial infection probability is fixed at ($\rho = 0.01$). Rows correspond to different probing budgets (k), while columns correspond to the sampling scheme: uniform sampling in the first column and weighted sampling in the second column. 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.

Author Biographies

Yeganeh Alimohammadi is an assistant professor of data sciences and operations at the University of Southern California's Marshall School of Business. Her research spans applied probability, network theory, and the design of markets and platforms.

Christian Borgs is a professor in the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley, a member of the Berkeley AI Research Lab, and faculty director of the Bakar Institute of Digital Materials for the Planet. He co-founded Microsoft Research New England. His work centers on the science of networks, including the theory of graph limits.

Remco van der Hofstad is a professor and chair of probability in the Department of Mathematics and Computer Science at Eindhoven University of Technology. His research focuses on probability theory, random graphs, and complex networks, and he is the author of the textbook series *Random Graphs and Complex Networks*.

Amin Saberi is a professor of management science and engineering, and by courtesy of computer science, at Stanford University, where he directs the Stanford Center for Computational Market Design. His research interests include algorithms, the design and analysis of networks, and their applications.

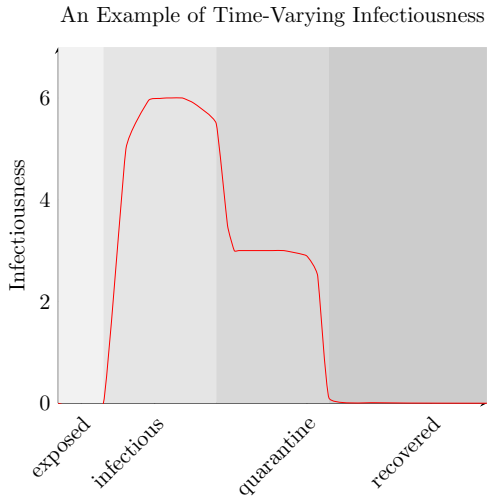

Figure 1 The density of infectiousness over time, illustrating the intervals of exposure, infectiousness, quarantine with reduced transmission rate, and recovery.

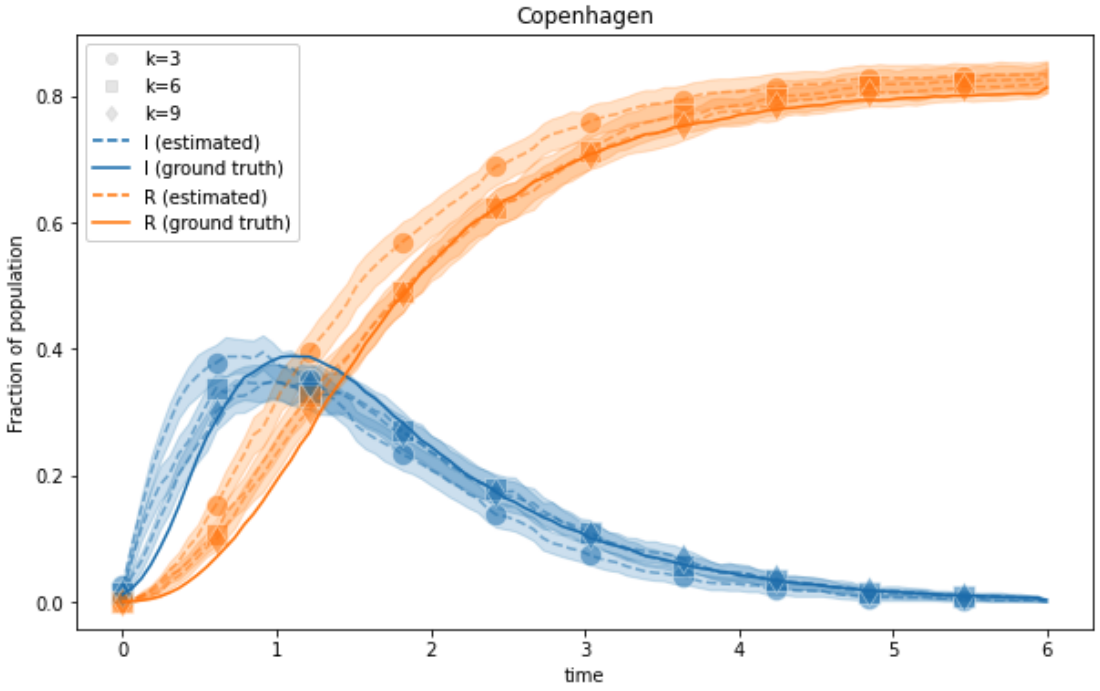

Figure 2 Copenhagen Dataset: Time Evolution of Epidemic Infections (I) and Recoveries (R) with Estimator Evolution for Various Testing Budgets (k)

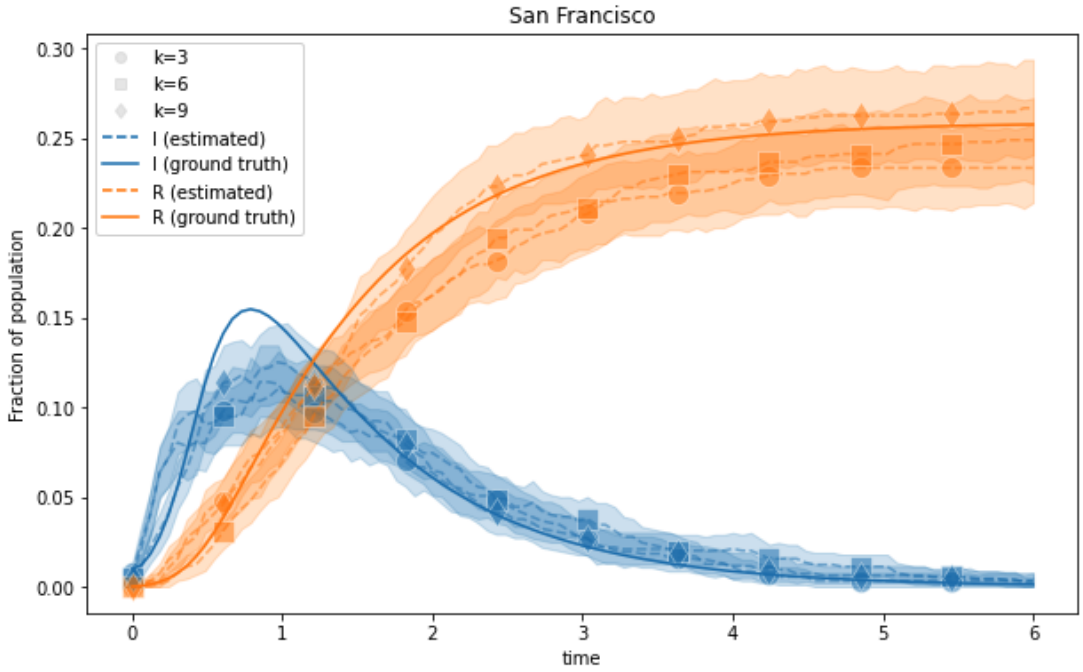

Figure 3 San Francisco Dataset: Time Evolution of Epidemic Infections (I) and Recoveries (R) with Estimator Evolution for Various Testing Budgets (k)

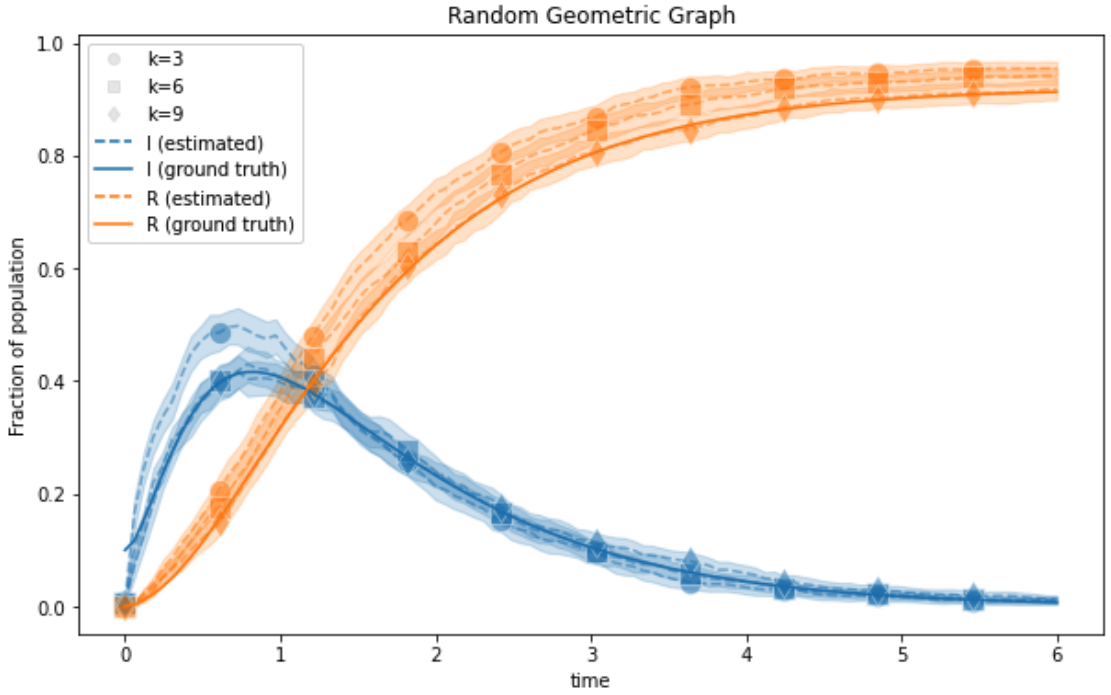

Figure 4 Random Geometric Graph: Time Evolution of Epidemic Infections (I) and Recoveries (R) with Estimator Evolution for Various Testing Budgets (k)

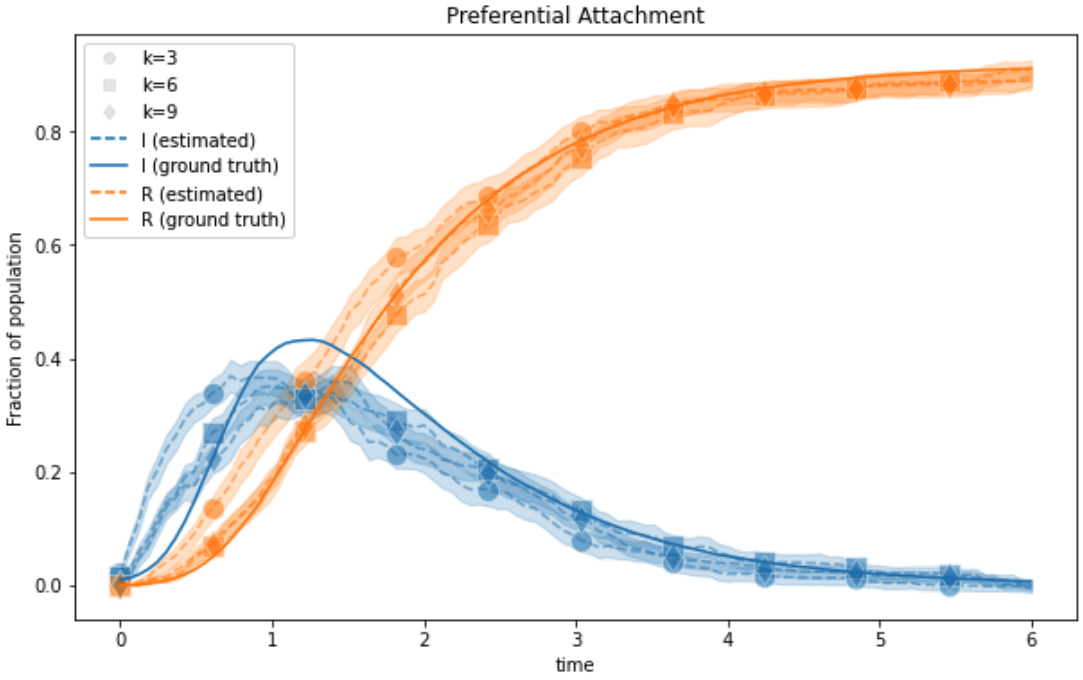

Figure 5 Preferential Attachment: Time Evolution of Epidemic Infections (I) and Recoveries (R) with Estimator Evolution for Various Testing Budgets (k)

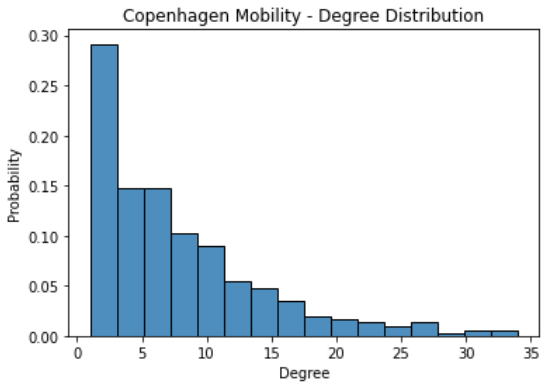

Figure 6 Degree Distribution of Copenhagen Network.

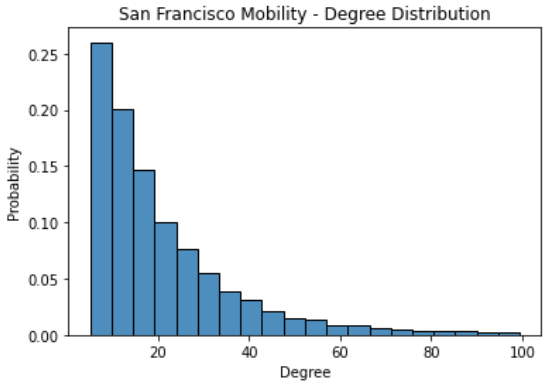

Figure 7 Degree Distribution of San Francisco Mobility Network (Capped at 100 for Depiction).

Table 1 Performance evaluation of the estimator on Copenhagen Dataset. 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	0.091 (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 Absolute Error (CI)	Time Evolution Euclidean Distance (CI)	Time Evolution 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 Absolute Error (CI)	Random Geometric Graph Absolute Error (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: $\hat{R}_{q,k,n}(\infty)$ (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 Fraction ρ_V	Final Size Abs. Error (CI)	Time Evolution Euclidean Distance (CI)	Time Evolution 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 Fraction ρ_V	Final Size Abs. Error (CI)	Time Evolution Euclidean Distance (CI)	Time Evolution 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)

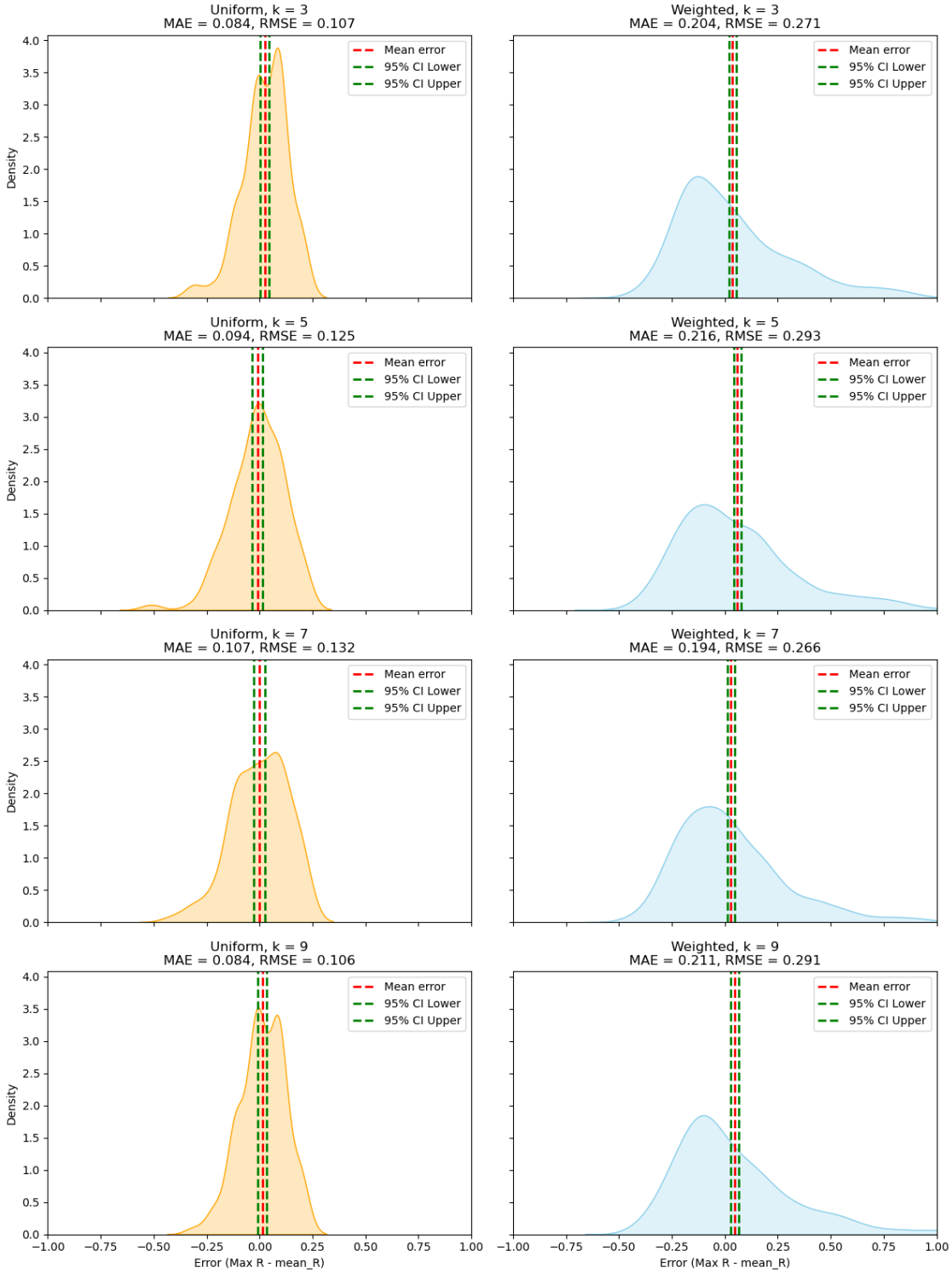

Figure 8 Error distributions for the final epidemic size $R(\infty)$ in **Copenhagen dataset** using two sampling strategies: uniform sampling (first column) and weighted sampling (second column, with selection probabilities proportional to node degree plus 0.01). Each plot shows the error distribution 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).

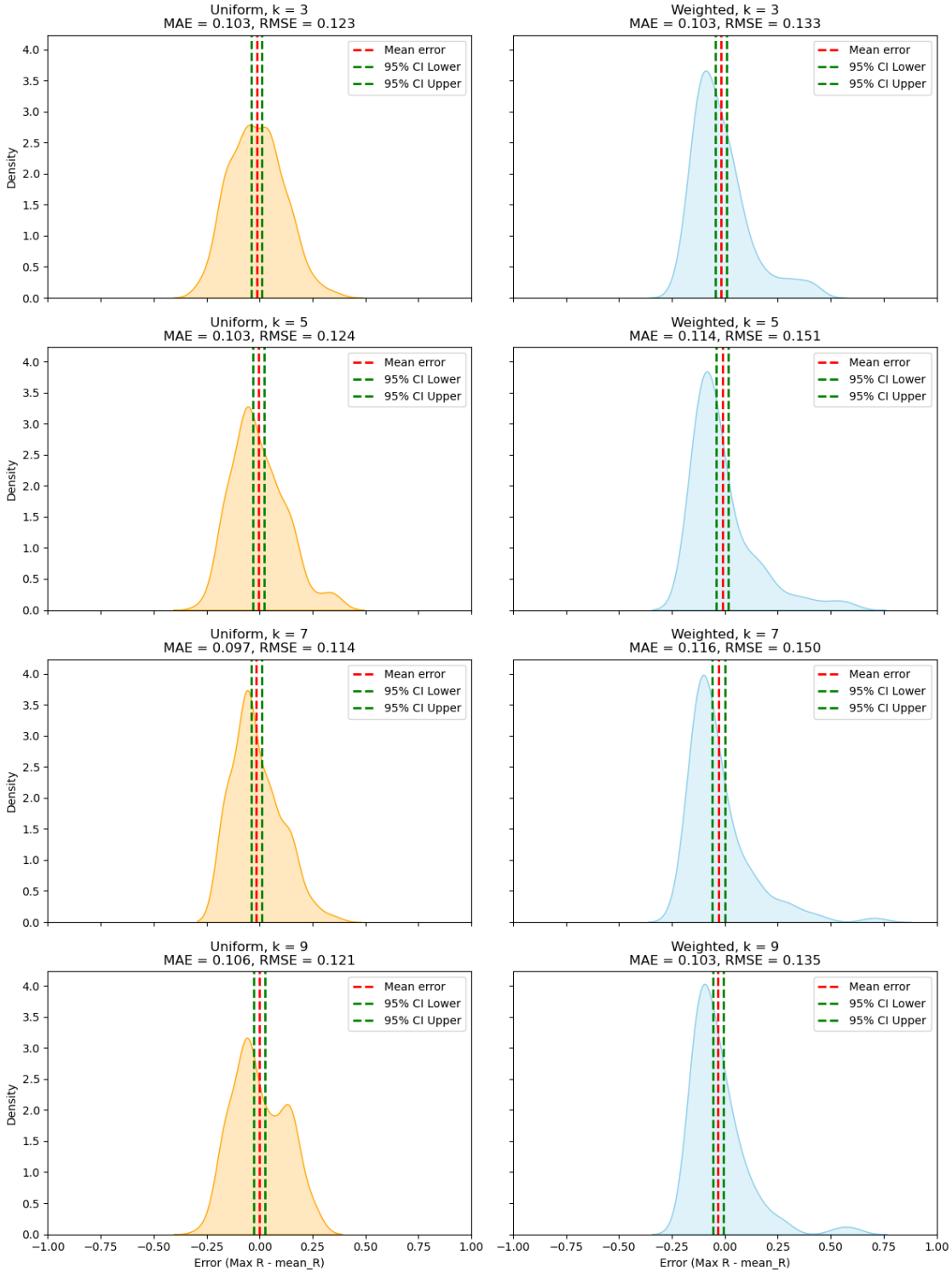

Figure 9 Error distributions for the final epidemic size $R(\infty)$ in **San Francisco dataset** using two sampling strategies: uniform sampling (first column) and weighted sampling (second column, with selection probabilities proportional to node degree plus 0.01). Each plot shows the error distribution 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).