SUPPLEMENT TO “NON-ROBUSTNESS OF DIFFUSION ESTIMATES ON NETWORKS WITH MEASUREMENT ERROR”

ARUN G. CHANDRASEKHAR, PAUL GOLDSMITH-PINKHAM, TYLER H. MCCORMICK, SAMUEL THAU, AND ZEYU WEI

1. Simulations

We simulate SIR diffusion on synthetic networks to illustrate the finite-sample behavior of our results.

We generate L_n by placing $n = 4{,}000$ nodes on a q-dimensional lattice in $[0,1]^q$ with additional uniformly placed nodes, linked to nearby nodes with a radius ensuring connectivity. We simulate two networks ($q = 4$ and $q = 2$) with $\mathcal{R}_0 = 2.5$ and $p_n = \mathcal{R}_0/\bar{d}$. Summary statistics appear in Table 1. We set T to twice the diameter of L_n ($T = 36$ for $q = 4$; $T = 186$ for $q = 2$), which extends beyond the intermediate-horizon regime of Assumption 1 into the saturation phase.[1]

Sensitive Dependence. We fix L_n, draw a single E_n, set i_0 at the lattice center, and construct the alternate seed set J_{i_0} at distance 2 from i_0. For $q = 4$, this neighborhood covers 1.85% of nodes; for $q = 2$, the neighborhood covers 0.45% of nodes in the graph. We approximate the Jaccard index from Theorem 1, $\mathcal{J}(i_0, j_0)$, by fixing E_n and averaging the Jaccard index $\mathcal{J}$ over 2,500 percolation draws.

Figures 1a and 1c show little overlap between diffusions until the network saturates. For $q = 4$ at $T = 10$ (half the diameter of L_n), $\mathcal{J} = 0.29$—nearly disjoint processes; under G_n, this value is $\mathcal{J} = 0.27$. For $q = 2$ at $T = 47$ (half the diameter of L_n), $\mathcal{J} = 0.75$ under L_n and 0.85 under G_n. Comparing dimensions: lower q produces more sensitivity in the *extent* of diffusion (i.i.d. connections generate more activations), while higher q produces more sensitivity in *location*. Both show severe sensitive dependence early on.

Forecast Errors. We draw E_n as an Erdős–Rényi graph with $\beta_n = 1/(10n)$, re-drawing E_n each of 2,500 iterations. The average additional degree from E_n is just 0.100, but the effect on global geometry is dramatic: for $q = 2$, adding E_n reduces average path length from 31.8 to 10.3 while mean degree rises only from 5.8 to 5.9—consistent with the path-length compression first observed numerically by

@@, and predicted by our framework for any polynomial-expansion base network with unaligned missing links. We fix a random seed i_0 and compute $\hat{Y}_T(L_n)/\hat{Y}_T(G_n)$.

Figures 1b and 1d show $\hat{Y}_T(L_n)/\hat{Y}_T(G_n)$ over time. For $q = 4$, the minimum ratio is 0.775 at $T = 13$ (a 22% underestimate); for $q = 2$, it drops to 0.164 at $T = 28$. Lower dimension produces greater sensitivity to additional links. Even with $\beta_n = 1/(100n)$ and $q = 2$, the minimum ratio remains far below the $q = 4$ values (Section 1.7). The curves match the theoretical prediction: the ratio declines until diffusion reaches

Department of Economics, Stanford University; NBER; J-PAL.

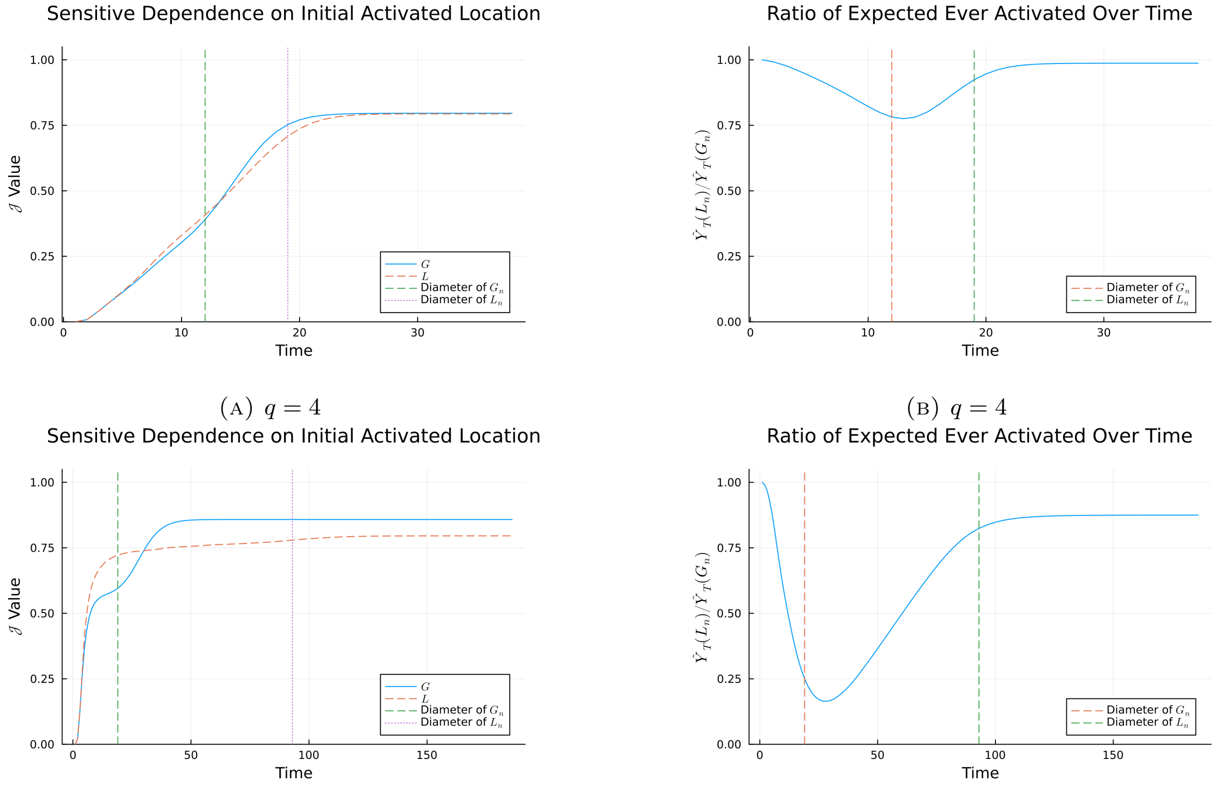
Yale School of Management; J-PAL.

Department of Statistics, University of Washington.

Department of Economics, Stanford University.

Department of Statistics, University of Washington.

[1] Recall the time period bounds from Assumption 1. We use horizons beyond $\overline{T}_n$ to show that the qualitative patterns predicted by the theory—forecast underestimation during the intermediate regime, followed by convergence as the network saturates—persist in finite samples even when the formal asymptotic bounds are not binding.



(A) $q = 4$

(B) $q = 4$

(C) $q = 2$

(D) $q = 2$

FIGURE 1. Panels 1a and 1c show simulations of Theorem 1, while panels 1b and 1d show simulations of Theorem 2. Panels 1a and 1c each fix a separate draw of E_n, then each choose a fixed j_0. We then simulate 2,500 diffusion processes while tracking the Jaccard index after perturbing the initial seed location. In Panels 1b and 1d, we simulate 2,500 iterations of the diffusion process on both L_n and G_n for each value of q, re-drawing E_n for each simulation. We then track the expected number of ever-activated nodes under each simulation at each time period, and then take the ratio.

the diameter of G_n, then recovers toward one as the network saturates.[2] Figure 2 separates the curves for $\hat{Y}_T(L_n)$ and $\hat{Y}_T(G_n)$, confirming that divergence peaks near the diameter of G_n.

1.1. Simulation Details.

We now describe the simulation procedures in detail.

1.2. Graph Generation.

Graph geometry plays a key role in our results. We build a network as follows, to generate an empirical analogue to the L_n that we study theoretically. L_n is generated as a graph of n nodes in the following manner.

(1) The base construction of the graph is a q-dimensional lattice, to mimic the properties of Assumption 2. We place n_{side} nodes evenly spaced on $[0, 1]^q$, meaning that there are n_{side}^q nodes in the lattice portion of the graph.
(2) The remainder of n nodes are placed uniformly at random throughout $[0, 1]^q$.

[2] The ratio asymptotes just below 1, since G_n permits slightly more activations in expectation than L_n.

(3) All nodes, regardless of whether they are in the lattice or placed randomly, link to all nodes within distance r. We set r as:

$$r = \max \left\{ \frac{1}{n_{side} - 1}, \frac{\sqrt{q}}{2} \frac{1}{n_{side} - 1} \right\}$$

This ensures that the graph is connected, even when q is large and thus nodes can be far apart.

We use the following parameters to generate L_n in the graphs used in the main texts. In the first specification, we set $n = 4,000$, $q = 4$ and $n_{side} = 7$. In the second specification, we set $n = 4,000$, $q = 2$, and $n_{side} = 50$. To generate G_n, we add links with i.i.d. probability β_n. As a base rate, we use $\beta_n = \frac{1}{10n}$ – in one variant of parameters, we set $\beta_n = \frac{1}{100n}$. Summary statistics are shown in Table 1, and for additional simulations in Table 3.

Table 1. Graph statistics for L_n with $n = 4,000$ nodes

Statistic	L_n	G_n	L_n	G_n
Dimension	4.0	4.0	2.0	2.0
Diameter	19.0	11.621	93.0	20.438
Mean Degree	10.164	10.263	5.826	5.926
Min Degree	3.0	3.091	2.0	2.0
Max Degree	24.0	24.107	16.0	16.126
Mean Clustering Coefficient	0.265	0.258	0.382	0.37
Average Path Length	7.592	6.017	31.942	10.304

For $q = 4$, 60 percent of nodes are in the lattice, while with $q = 2$ 62.5 percent are. Statistics for G_n are the expectation over 2,500 draws of E_n, which is drawn Erdős–Rényi with $n = 4,000$ and $\beta_n = \frac{1}{10n} = \frac{1}{40000}$.

1.3. **Diffusion Process.** We use a Susceptible-Infected-Removed (SIR) diffusion process. Each node is infected (activated) for a single period, and has the opportunity to transmit the process with i.i.d. probability p_n to each of its neighbors. After nodes are activated, they are removed and cannot be re-activated. We set the basic reproductive number to be $\mathcal{R}_0 = 2.5$, and set $p_n = \mathcal{R}_0/\bar{d}$, where $\bar{d}$ is the mean degree in L_n.

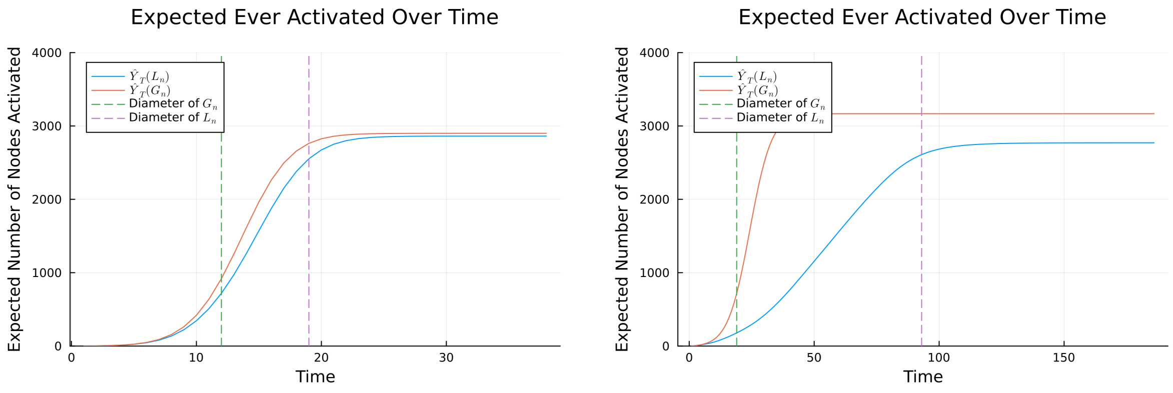
1.4. **Simulation of Theorem 1.** As an analogue to Theorem 1, we simulate SIR processes on a fixed G_n with slightly perturbed starting points. We choose i_0 to be in the center of the lattice of L_n, that forms the backbone of G_n. Then, we build a set of alternative seeds J_{i_0}. All nodes at 2 are included in J_{i_0}. We then choose a $j_0 \in J_{i_0}$ uniformly at random.

The SIR process is then run, starting at both i_0 and j_0. We record which nodes are ever activated at each step of the process, under each simulation. To follow Theorem 1, we fix the percolation across the simulation starting at i_0 and j_0. To do so, we use the fact that for a one-period SIR model, each link can transmit the disease at most one time. Therefore, we can simulate ex-ante which links will be able to transmit, which occurs with probability p_n, and intersect this with G_n to get the realized percolation.

We then compute a standard Jaccard index to track the intersection of the two epidemics. Let $I_P(i_0)$ be the set of ever-activated nodes under the epidemic from i_0, and $I_P(j_0)$ be the corresponding set from j_0. Then, we compute:

$$\mathcal{J} := \mathbb{E}_P \left[\frac{|I_P(i_0) \cap I_P(j_0)|}{|I_P(i_0) \cup I_P(j_0)|} \mid L_n, E_n \right]$$

We work with the expected Jaccard index, rather than consider the probability the Jaccard index is bounded away from one.



(A) $q = 4$

(B) $q = 2$

FIGURE 2. This figure plots the same information as Figure 1, but separated by graph for both $q = 4$ and $q = 2$. The trajectory of $\hat{Y}_T(L_n)$ initially lags behind that of $\hat{Y}_T(G_n)$, leading to the decrease in the ratio shown in Figure 1. As $\hat{Y}_T(L_n)$ catches up, the ratio increases.

1.5. Simulation of Theorem 2.

To investigate the content of Theorem 2, we directly simulate the sample analogue. For 2,500 simulations, we do the following. We choose the initial seed i_0 uniformly at random, and fix it throughout the process. The SIR process is simulated for T periods, where we set T to be twice the diameter of L_n.

(1) Simulate the SIR process on L_n.
(2) Generate a draw of E_n, with links i.i.d. with probability β_n.
(3) We define $G_n := L_n \cup E_n$, and simulate the SIR process on G_n.

We track the number of ever-activated nodes in each simulation at each time step. We then take the average over simulations at each time step. Results are shown in Figure 1. Additional results are shown in Figures 2 and 3.

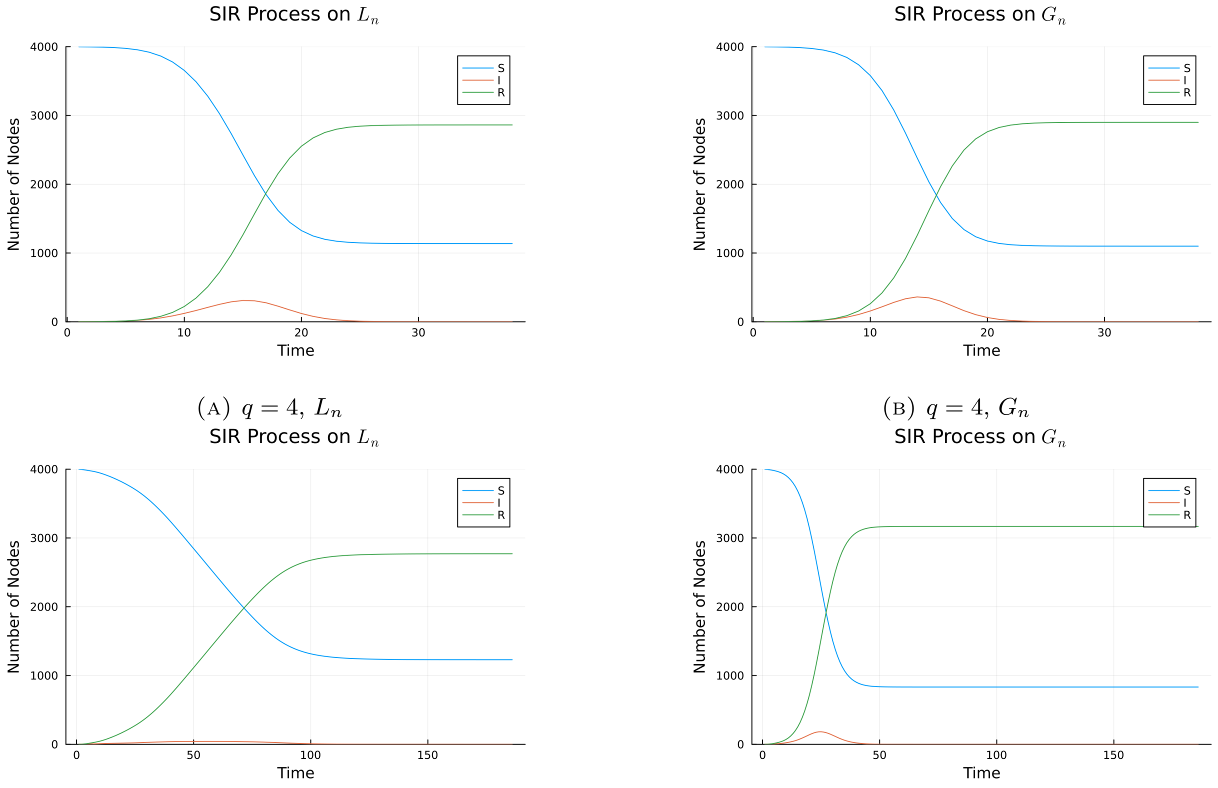
1.6. Simulation of Theorem 3 (Aligned Error).

We simulate a version of Theorem 3. We begin with the base graph (what we used as L_n in the prior two sections) and remove links with i.i.d. probability—producing aligned error, since missing links are a uniform thinning of the true network. Here, we choose the deletion probability ε_n to delete the same number of links on average as we did when we added links with probability $\beta_n = \frac{1}{10n}$. We then investigate the objectives from Theorems 1 and 2, taking the base graph as G_n and the thinned graph as L_n. Results are shown in Figure 4. We can see the system exhibits minimal additional sensitive dependence on the seed in both the $q = 4$ and $q = 2$ cases, beyond the base level of sensitivity from prior simulations. We also see reduced underestimation of the diffusion ratio $\hat{Y}(L_n)/\hat{Y}(G_n)$: for $q = 4$, the minimal value is 0.96 at $T = 10$, while for $q = 2$ it is 0.89 at $T = 44$.

1.7. Extreme Sensitivity with $q = 2$.

We explore an additional set of simulations in the case of $q = 2$, this time using a much smaller value of $\beta_n = \frac{1}{100n}$. We show average graph statistics in Table 3. Results are shown in Figures 5.

As shown in Figure 5, despite a much smaller value of β_n forecasting issues persist. There remains meaningful sensitive dependence on the initial seed—the average $\mathcal{J}$ at half the diameter of L_n is 0.74 on L_n and 0.75 on G_n, indicating that roughly 26% of the activation set changes when the seed is perturbed by two hops. For forecasting diffusion volume, the minimum value of $\hat{Y}_T(L_n)/\hat{Y}_T(G_n)$ is achieved at $T = 47$,



(A) $q = 4$, L_n

(B) $q = 4$, G_n

(C) $q = 2$, L_n

(D) $q = 2$, G_n

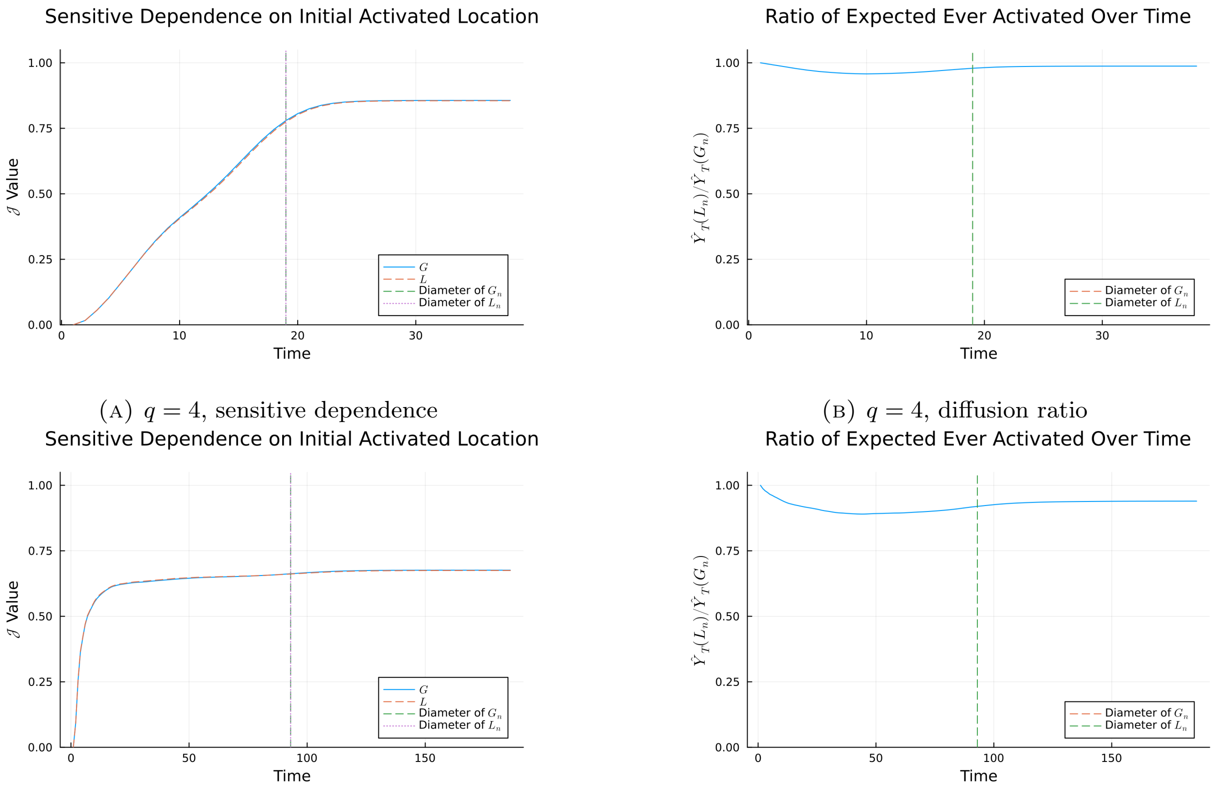
FIGURE 3. Simulations meant to emulate Theorem 2, disaggregated into the standard SIR framework. The figure is a result of averaging over simulation draws. Note that we see a larger spike in activations under G_n, which makes intuitive sense – the additional links allow for more infections to occur. We show results for both $q = 4$ and $q = 2$, both with $\beta_n = \frac{1}{10n}$. Note that the gap between total activations with $q = 2$ is larger, as the additional links have a larger effect.

TABLE 2. Graph statistics for aligned-error simulations with $n = 4,000$ nodes

Statistic	L_n	G_n	L_n	G_n
Dimension	4.0	4.0	2.0	2.0
Diameter	18.805	19.0	92.858	93.0
Mean Degree	10.064	10.164	5.727	5.826
Min Degree	2.951	3.0	1.49	2.0
Max Degree	23.783	24.0	15.758	16.0
Mean Clustering Coefficient	0.261	0.269	0.375	0.383
Average Path Length	7.584	7.562	31.938	31.806

L_n is generated from G_n by thinning the network, removing links with i.i.d. probability ε_n. We calibrate ε_n so that the expected difference in link volume between L_n and G_n is the same as in simulations where we generate G_n by adding links to L_n. The diameter is computed over the largest connected component of the graph – in some situations, the diameter decreases due to the thinning as pieces of the graph become disconnected.

taking a value of 0.64. This value is still lower than the case with $q = 4$ and $\beta_n = \frac{1}{10n}$, showing the extreme sensitivity in the lower dimension.



(A) $q = 4$, sensitive dependence

(B) $q = 4$, diffusion ratio

(C) $q = 2$, sensitive dependence

(D) $q = 2$, diffusion ratio

FIGURE 4. Simulations meant to emulate Theorem 3, using the objectives from Theorems 1 and 2. We show results for both $q = 4$ and $q = 2$, in both cases calibrating the probability of link deletion so that the same volume of links are deleted as in the case where we add links with probability $\frac{1}{10n}$.

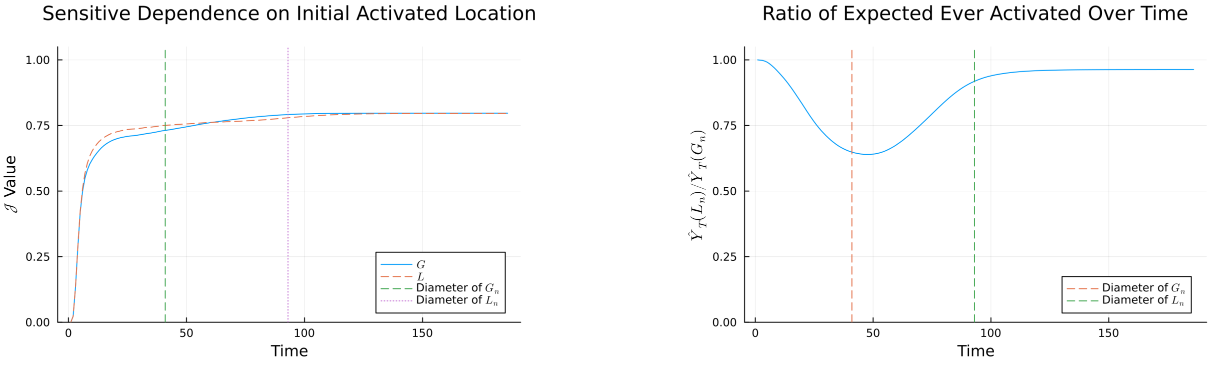
TABLE 3. Graph statistics for L_n generated with $q = 2$ and G_n generated with $\beta_n = \frac{1}{100n}$

Statistic	L_n	G_n
Dimension	2.0	2.0
Diameter	93.0	45.059
Mean Degree	5.826	5.836
Min Degree	2.0	2.0
Max Degree	16.0	16.007
Mean Clustering Coefficient	0.379	0.38
Average Path Length	31.774	18.802

Statistics for G_n are taken as an average over 2,500 draws.

2. Additional Results for Location Data from the COVID-19 Epidemic

2.1. **Lower i.i.d. error rates.** To make a more direct comparison to the Monte Carlo simulations, we repeat the simulation exercises using E_n generated i.i.d. with $\beta_n = \frac{1}{10n}$. Graph statistics are shown in Table 4, again for L_n and the average statistics for G_n over 2,500 draws of E_n. Compared to G_n in the main text (in Table F.1), note that the change in degree, clustering, and average path length are all much smaller, as E_n is much more sparse in this case.



(A)

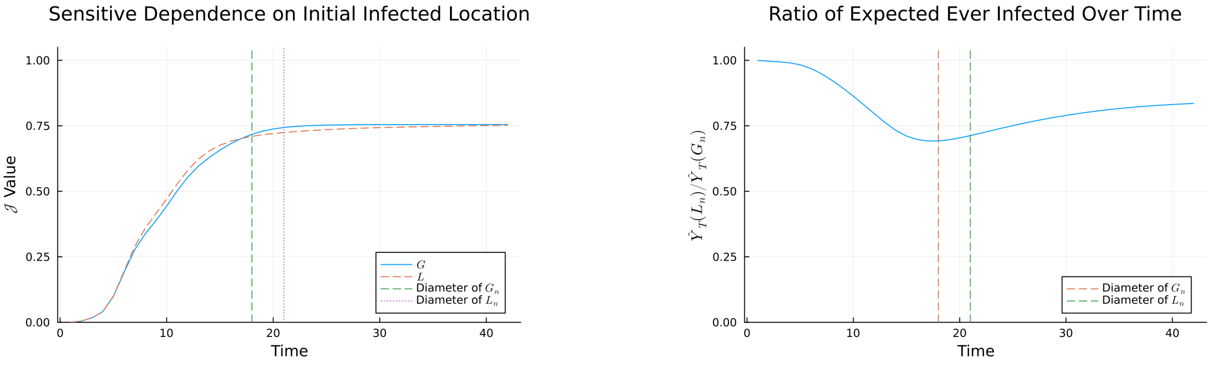
(B)

FIGURE 5. Results with $q = 2$ and $\beta_n = \frac{1}{100n}$. Panel (A) shows the Jaccard index $\mathcal{J}$, while Panel (B) shows the ratio $\hat{Y}_T(L_n)/\hat{Y}_T(G_n)$. Averages are taken over 2,500 Monte Carlo simulations.

TABLE 4. Average graph statistics with i.i.d. errors in the travel data for California, Nevada, and a small portion of Arizona

Statistic	L_n	G_n
Diameter	21.0	16.914
Mean Degree	12.962	13.062
Min Degree	1.0	1.0
Max Degree	298.0	298.106
Mean Clustering Coefficient	0.39	0.381
Average Path Length	7.287	6.117

G_n is generated from L_n using i.i.d. additional links, which occur with $\beta_n = \frac{1}{10n}$.



(A)

(B)

FIGURE 6. Results using the COVID-19 travel data, with G_n using E_n generated i.i.d. with $\beta_n = \frac{1}{10n}$. Panel (A) shows the ratio $\hat{Y}_T(L_n)/\hat{Y}_T(G_n)$, while Panel (B) shows the Jaccard index $\mathcal{J}$. Averages are taken over 2,500 Monte Carlo simulations.

Results are shown in Figure 6. We take averages over 2,500 simulations. The top left panel shows the simulation of Theorem 1. As in the main text, we choose the local neighborhood containing all j_0 conservatively: we chose the set to be all nodes within distance 2 of i_0. The distance from i_0 to j_0 is therefore 2, and the neighborhood that contains all possible j_0 contains 0.80 percent of the graph. Halfway

Table 5. Average graph statistics with aligned errors in the travel data for California, Nevada, and a small portion of Arizona

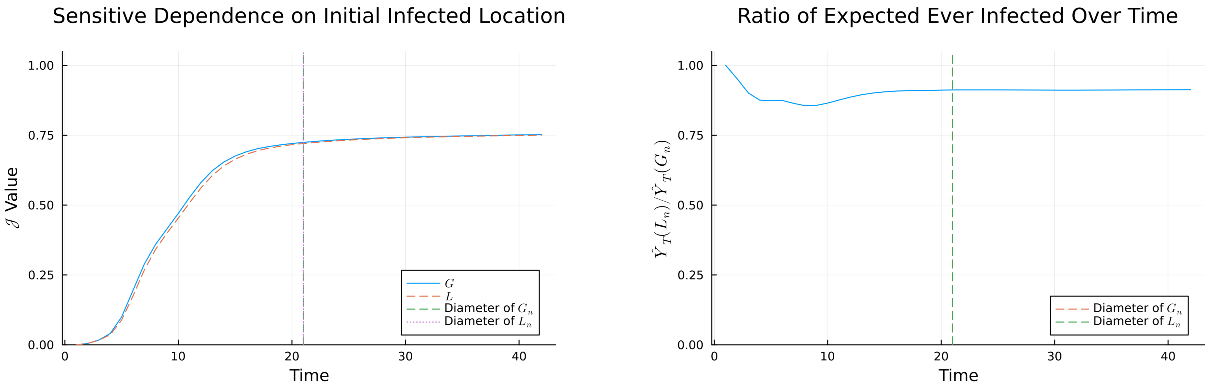
Statistic	L_n	G_n
Diameter	21.388	21.0
Mean Degree	12.126	12.962
Min Degree	0.156	1.0
Max Degree	278.878	298.0
Mean Clustering Coefficient	0.363	0.388
Average Path Length	7.479	7.277

L_n is generated from G_n dropping links with probability ε_n, i.i.d., with $\varepsilon_n = \frac{1}{\mathrm{diam}(G_n)^3}$. Note that G_n corresponds to the base graph, L_n, in the i.i.d. and pruned cases. Values for L_n are averages taken over 2,500 draws.

to the diameter of L_n, the value of the average Jaccard index is 0.47 under L_n and 0.44 under G_n, indicating largely distinct epidemics. The top right panel shows the simulation of Theorem 2. Note that in this case, the minimum ratio of $\hat{Y}_T(L_n)/\hat{Y}_T(G_n)$ is achieved at $T = 17$ and takes the value 0.69. This value is much larger than the values from the main text with either the pruned or i.i.d. errors, and comparable to the values with the same level of β_n and graph dimension $q = 4$ in the Monte Carlo simulations.

2.2. **Lower aligned error rates.** Here, we consider aligned error, but with a lower rate of thinning. Specifically, we set $\varepsilon_n = \frac{1}{\mathrm{diam}(G_n)^3}$, to match Theorem 3. Summary statistics for the resulting graph are shown in Table 5. Note that G_n corresponds to L_n in the pruned and i.i.d. error simulations.

Results are shown in Figure 7. Here, the gap between the Jaccard indices at time step 10 is 0.017, at 0.47 and 0.45. The minimum ratio $\hat{Y}(L_n)/\hat{Y}(G_n)$ is 0.85, at $T = 8$. This minimum is much larger than in the other cases.



(a) Sensitive dependence under aligned error

(b) Forecast error under aligned error

Figure 7. Figures show the average Jaccard index for diffusion starting at i_0 and a specific j_0 and diffusion ratio $\hat{Y}_T(L_n)/\hat{Y}_T(G_n)$ under aligned errors with a lower rate of missing links.