# NON-ROBUSTNESS OF DIFFUSION ESTIMATES ON NETWORKS WITH MEASUREMENT ERROR

ARUN G. CHANDRASEKHAR[‡,†,⋆], PAUL GOLDSMITH-PINKHAM[∘,⋆], TYLER H. MCCORMICK[§], SAMUEL THAU[‡], AND ZEYU WEI[§]

Abstract. Network diffusion models are used to study disease transmission, information spread, technology adoption, and other socio-economic processes. We show that estimates of these diffusions are highly non-robust to mismeasurement. First, even when the *network* is measured perfectly, small and local mismeasurement in the initial seed generates a large shift in the locations of the expected diffusion. Second, if instead the initial seed is known, small and arbitrarily structured measurement error in links, with the share of missed links close to zero, causes diffusion forecasts to be significant under-estimates. Such failures exist even when the basic reproductive number is consistently estimable. We explore difficulties implementing possible solutions, such as estimating the measurement error or implementing widespread detection efforts. Finally, we conduct simulations on synthetic and real networks from three settings: travel data from the COVID-19 pandemic, a mobile phone marketing campaign in rural India, and an insurance experiment in China.

Researchers and policymakers studying the spread of ideas, technology, or disease often use network data on how individuals interact to estimate models of diffusion. Examples include (i) quantifying the extent of illness or technology take-up; (ii) summarizing diffusion dynamics (e.g., the reproduction number $\mathcal{R}_0$ of a disease); (iii) targeting interventions (e.g., where to seed new information to maximize spread, where to lockdown to prevent spread); (iv) and estimating counterfactuals (e.g., in estimates of peer effects, as we show in an empirical example). See Anderson and May (1991), Jackson (2009), Jackson and Yariv (2011), and Sadler (2023) and references within for all three classes of topics (as well as an account of how such models are used in the case of strategic behavior).

We study when the econometrician has imperfect measurement of either the initial seeding or the interaction network, and wants to estimate models of diffusion or generate diffusion forecasts. Importantly, we consider cases where the measurement error (1) is very small, such that it vanishes in the limit of our asymptotic case, and (2) only exists in either the initial seed or underlying network, giving perfect knowledge to the econometrician about the network or seed, respectively. Nonetheless, we show that tiny mismeasurement significantly affects the predictions of the econometrician's estimated diffusion model.

We focus on an intermediate time horizon, such that the econometrician is not focused on predictions on "day 2" of a diffusion, nor are they focused on "long-run" predictions. The short-term is highly stochastic, while, in the long-run, the diffusion will completely saturate the network and forecasts and predictions of where the diffusion goes are much less consequential for policy. Our preferred environment captures the

We gratefully acknowledge Isaiah Andrews, Abhijit Banerjee, Haoge Chang, Jishnu Das, Matt Jackson, Matt Gentzkow, Ben Golub, Ed Kaplan, Julianne Meisner, Jim Moody, Karl Rohe, Adam Szeidl, and Juan Pablo Xandri.

[‡] Department of Economics, Stanford University.

[†] NBER.

[⋆] J-PAL.

[∘] Yale School of Management.

[§] Department of Statistics, University of Washington.

[¶] Department of Sociology, University of Washington.

setting where an econometrician is equipped with the richest possible data on individuals and interactions, and wants to predict the diffusion process in the policy-relevant "medium-run."

We show four key results: (i) predictions of where diffusion goes is sensitive to *local* uncertainty of the initial seeding; (ii) predictions of diffusion counts will be under-estimated with even vanishingly small measurement error of the network; (iii) while aggregated estimated quantities such as the basic reproductive number $\mathcal{R}_0$ can be estimated correctly despite the measurement error, it provides limited information for more disaggregated targets; (iv) because the measurement error is so small, most data augmentation (either estimating the measurement error or conducting additional data collection) will be ineffectual. These seemingly pessimistic results, however, provide clarity on possible positive strategies to explore: the extremely high returns to wide-spread strategies early in a diffusion.

To give intuition, consider a network formation process where connections occur with a higher probability for people with some observable commonality (e.g., geography, school, work) or latent factors (e.g., Hoff et al. (2002)). With a perfectly measured graph, when a diffusion process is seeded, we can draw a ball around the (known) initial seed that will exhaustively enumerate the number of nodes possibly affected by the process. This ball will expand over time, with the ball's radius defined by the distance from the initial seed. Even if the network is fully known, proximate initial seeds can effectively have the balls drawn around them expand differently—there will be overlap between the expanding shells, but also non-trivial divergence. So, small perturbations in the initial seed can lead to misleading conclusions on where the diffusion process goes.

Now, in contrast, imagine the seed is known, but a small set of idiosyncratic links are missed and unobserved. If any of these missed links reach further than the ball drawn around the seed in the observed graph, the diffusion process will escape past the econometricians' determined set of possibly impacted nodes. Once the link is outside of the ball, it spreads even more quickly because it can diffuse to the largest possible set of unexposed units. This jump need not be far – it simply needs to be a link that creates diffusion unexpected by the econometrician.[1] In our general theorems, each node can link to a (vanishing) fraction of the population and the linking can be arbitrary in structure. This nests cases of only local mismeasurement: e.g., only missing links to "nearby" locations.

Missing links in the measurement of networks is a common concern (Wang et al., 2012; Sojourner, 2013; Chandrasekhar and Lewis, 2010; Advani and Malde, 2018; Griffith, 2022), but our paper highlights the dramatic impact of even the smallest errors when forecasting diffusion. Mismeasurement can happen for several reasons. The first is practical: many analyses using empirical data (including one of our own empirical examples) do some amount of aggregation into groups with measured amounts of interaction. For example, individuals may be binned into groups of location-by-age-by-occupation, and the interactions between these groups are approximated based on underlying microdata. Using these data on individuals and interactions to construct compartments and forecast diffusion processes implicitly assumes that connections occur with a much higher probability for people with some observable commonality within the bin (Acemoglu et al., 2021; Farboodi et al., 2021; Fajgelbaum et al., 2021). These choices may match the average interaction pattern,

[1] Our analysis is related to but distinct from Watts and Strogatz (1998). First and foremost, we do not require that the missed links could go anywhere in the network. Our most general results allow for nodes to have mismeasurement to potentially only a vanishing share of nodes in the graph. In our environment, the key condition of polynomial expansion is a joint property of the graph and diffusion process and not a property of the graph alone. This distinct assumption allows for analytic analysis of the diffusion processes, while also allowing for a much wider array of graph structures (including expansive networks). Further, much of the work on small world graphs and diffusion focuses on phase transitions of the process (e.g. Newman and Watts (1999)), but we compare shifts within the same (critical) phase. Finally, our focus is on forecasts of the extent and location of the diffusion, sensitivities to perturbation of the initial seed, and possible solutions to the identified problems.

but miss underlying heterogeneity, and may also mismeasure cross-compartment connections. Second, the mismeasurement of the network may occur because the sampling process for the network is imperfect. Studies surveying individuals may focus on local connections (e.g. within a school or village) and ignore other connections. Or, it may be that certain connections are not mentioned, despite mattering to the diffusion process. Third, it may be that a rich snapshot of a network does not capture the relevant links for diffusion by the time the process reaches an individual.

Formally, we study an asymptotic model in which parameters depend on the number of agents to tractably approximate finite-sample/time behavior, as is customary in graph theory. We consider a triangular array where there is a discrete set of $n$ agents who are in an undirected, unweighted network $G_n$. We take $n \to \infty$. A SIR (susceptible-infected-recovered) diffusion process proceeds for $T_n$ periods on $G_n$. In each period, a newly infected node passes the disease i.i.d. with probability $p_n$ to each of its neighbors and is then removed from the process. Since the model applies to diseases, technology adoption, social learning, and other diffusion settings, we use the term *activated* to nest the application-specific terms such as "infected," "informed," or "adopted" (Jackson and Yariv, 2007).

We consider a time regime where it is neither early nor late. In a model that studies diffusion, this is the setting that is policy relevant. Early on, there is almost no information and nothing has happened. Similarly, if we look far into the future, then the diffusion process will have saturated the network. To approximate finite-time behavior, we impose that $T_n$ is an increasing function of $n$. Our precise conditions embed an upper bound on $T_n$, which ensures that the diffusion does not cover the entire graph, while a corresponding lower bound ensures that we consider cases beyond an extremely local forecasting problem. The resulting asymptotic framework applies to *any* $T_n$ that falls within the given bounds, which are determined from the structure of the model. This asymptotic framework is a convenient structure for considering diffusion processes in the "medium-run."

We define the true network over which the diffusion process spreads as $G_n = L_n \cup E_n$. The subgraph $L_n$ is fully observed by the econometrician and is deterministic, while $E_n$ is an unobserved stochastic error graph[2]. Motivated by the empirical and statistical literature, we assume that the diffusion process on $L_n$ has a predominantly polynomial expansion structure in our main results, which generalizes a local meeting topology (e.g., geography, social groups). The error graph $E_n$ contains idiosyncratic links that, in our baseline model, are drawn i.i.d. with probability $\beta_n$ among pairs of nodes. For each node, we allow it to link to a fraction $\delta_n$ of nodes; when $\delta_n = 1$, this corresponds to i.i.d. links between all pairs of nodes, but our results allow for $\delta_n$ to vanish. Crucially, we assume that in the limit asymptotically all the links in the true network $G_n$ are observed: $|E_n|/|L_n| \to_p 0$. We assume an even stronger upper bound on the rate so that no giant component can form in $E_n$: $\beta_n = o(1/n)$.

We proceed as follows. Sections 2 and 3 consider problems with forecasting. First, in Theorem 1, we show that diffusion on $G_n$ – *even when the error network is completely known* – is not stable with respect to the location of the initial activation. Theorem 2 shows that when the initial activation is known, but the error network is not known, the econometrician's estimates of the diffusion count will be of lower order of magnitude than the true counts in the intermediate run – the prediction will be dominated by the error. Section 4.1 demonstrates we can consistently estimate both the activation rate $p_n$ and the basic reproductive

[2] We focus on the case of missing links, as we believe this issue will be the primary one in practice (see Griffith (2022) for several empirical examples). In the case where the econometrician both misses some links and incorrectly assumes that others exist, the problem becomes much more complex. Globally, the net rate of missing or added links appears to be the key factor; locally, forecasts could over or underestimate the volume of diffusion

number $\mathcal{R}_0$, despite the aforementioned problems with forecasting diffusion[3]. In Section 4.2, we consider two possible solutions: (i) estimating the idiosyncratic links through supplementary data collection and (ii) widespread node-level sampling (e.g., testing). In our assumed regime neither solution works. The sample size required to estimate $\beta_n$ is unrealistically large, and the fraction of correctly identified locations with positive tests will be strictly below one in the short run.

We then examine versions of our main theorems on simulated networks. In Monte Carlo exercises, we generate networks that match known features of empirical data. We set the measurement error probability to be small ($\beta_n \approx 1/10n$) and find that underestimates of the diffusion count range from 22% to 83% across the simulations. We also demonstrate extreme sensitivity to initial conditions. When we perturb the initial seed in a neighborhood comprising 1% (or 5%) of the graph, the expected overlap share of activated nodes over perturbations is only 40% (or 13%) by the time the diffusion could potentially have saturated the network.

We then turn to analysis on real-world networks. In our first example, we construct a mobility network from the Southwest of the United States and examine mismeasurement due to "pruning" – where links between locations are only included if a sufficient number of people move between them. We find that changing the threshold from five to six people traveling between Census tracts causes the policymaker to underestimate the extent of diffusion by 56%. In addition to pruning, we induce errors by removing i.i.d. random links and find more extreme underestimation by more than 76%. As a second example, we show that similar patterns hold in a viral marketing experiment in rural India (Banerjee et al., 2019). We also document extreme sensitive dependence on the seed set: when we move only one single seed to one of its neighbors, the intersection is only 61% of the activations encompassed by both diffusions. Finally, we show how our results relate to the estimation of peer effects, focusing on the up-take of insurance in China (Cai et al., 2015).

## 1. Model

Environment. For a given set of observed nodes $V_n$ with the number of nodes $n$, we model the network through a random undirected and unweighted graph $G_n := (V_n, L_n \cup E_n)$ where $L_n$ consists of the "base" links and $E_n$ collects the missing links. Generally, we assume that $L_n$ is fixed and known perfectly, and all the links in $L_n$ are true links. Each link in $E_n$ is constructed independently following $\text{Ber}(\beta_{ij,n})$ where these can be heterogeneous at the pair level. The links in $E_n$ are random and not observed, and hence the randomness of the true graph $G_n$ only comes from the random realizations of $E_n$. Particularly, in our model we only consider the mismeasurement caused by missingness and there are no falsely added links. With an abuse of notation, we use $L_n$ and $E_n$ to denote the undirected and unweighted graphs with the base links and the missing links.

The diffusion process spreads over the network $G_n$ following a standard Susceptible-Infected-Removed (SIR) process with i.i.d. passing probability $p_n$. Each node is 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 for a single period, they are removed and cannot be re-activated. To better represent the randomness in the diffusion process, we define $P_n(G_n)$ as a random percolation on the graph $G_n$, which is a directed, binary graph with each directed link based on $G_n$ activated i.i.d. with probability $p_n$. The diffusion process is equivalent to a deterministic process emanating from some initial seed through $P_n(G_n)$, and hence

[3] Alimohammadi et al. (2023) makes a similar point. They study a SIR model on a network and design an estimation strategy for the parameters and the trajectory of epidemics. They consider a local estimation algorithm based on sampled network data, and show that asymptotically they identify the correct proportions of nodes that will eventually be in the SIR compartments. These results are analogous to our finding that one can estimate $p_n$ and $\mathcal{R}_0$ in a straightforward manner.

the randomness in the diffusion process is captured by the random realization of $P_n(G_n)$. Similarly, we let $P_n(L_n)$ as the percolation on the graph $L_n$ given by restricting $P_n(G_n)$ with the edges in $L_n$.

We conduct asymptotic analysis, taking limits as both $T$, the number of time periods, and $n$, the number of nodes, become large. We consider a sequence of graphs $\{G_n\} = \{L_n, E_n\}$, where $E_n$ are drawn randomly, that grows with $n$, and consider $T := T(n)$ where $T$ is an increasing function in $n$. More details on exactly how $T$ grows are discussed below. We generally suppress the dependence of $T$ on $n$ for simplicity.

Let $B_j(t)$ denote the ball of radius $t$ around vertex $j$ in a given graph and let $\{X_n\}_n = (X_1, \ldots, X_n, \ldots)$ denote the sequence of random variables. We define the expected activation for a diffusion on $L_n$ set as

$$\mathcal{E}_t = \mathbb{E}\left|\{x \in V_n \mid x \text{ ever activated by the diffusion on } L_n\}\right|.$$

To set up our first results, we impose the following condition on the diffusion process.

**Assumption 1.** *For some constant $q > 1$ and all discrete time $t$, $\mathcal{E}_t = \Theta(t^{q+1})$ and $\mathcal{S}_t = \mathcal{E}_t - \mathcal{E}_{t-1} = \Theta(t^q)$.*[4] *Furthermore, $p_n \in \left((\log n)^{-q/(2q+2)}, 1\right]$.* [5]

We write this assumption over the diffusion process rather than on the graph structure of $L_n$ to allow for more generality. We could have simply assumed that $L_n$ itself has polynomial expansion, and, together with the appropriate $p_n$ and i.i.d. draw assumptions, Assumption 1 follows. But, we allow for more general settings. For example, Assumption 1 covers cases of $L_n$ with non-polynomial expansion and i.i.d. draws of $p_n$, but with a sub-critical passing probability or short time horizons. The lower bound on $p_n$ is to ensure that the diffusion process spreads with sufficient speed – otherwise, the diffusion may halt before the medium time horizon that we study. We note three elements of substance on the assumption. First, it implies that as the diffusion progresses, a growing number of nodes become activated in expectation.[6] Second, this condition governs both the structure of the graph and the diffusion process. As an example, consider a latent space network where nodes form links locally in a Euclidean space (Hoff et al., 2002). Since volumes in Euclidean space expand at a polynomial rate, this ensures that Assumption 1 will be satisfied.[7] Third, note the geometric relationship between $\mathcal{E}_t$ and $\mathcal{S}_t$ — $\mathcal{E}_t$ governs the total volumetric expansion of the diffusion, while $\mathcal{S}_t$ governs the shells of the diffusion (e.g., the boundary at time $t$). We explore the case where $\mathcal{E}_t$ has exponential growth for completeness in Section 5.

We then put specific constraints on the time horizon considered. The first condition restricts the time so that the diffusion has not reached the edge of the graph.[8] The second condition ensures we are making a forecast about a time period appreciably far in the future. Let $a$ be any positive constant satisfying $2a > 1/(q+1)$; smaller $a$ are permitted for more expansive (larger $q$) graphs.

**Assumption 2.** *$T_n$ has for each $n$, $T_n \in [\underline{T}_n, \overline{T}_n]$ where the following holds: (1) $\overline{T}_n = n^{\frac{1}{q+1}}$ and (2) $\underline{T}_n = (\log n)^a$.*

Note that our main results hold for any $T_n \in [\underline{T}_n, \overline{T}_n]$. To get a sense of scale, consider California with a population of 39 million and assume the parameters are set at the day-level. If a geographic-type network

[4] $a_n \in \Theta(b_n)$ means $a_n$ is bounded both above and below by $b_n$ asymptotically in Bachmann-Landau notation.

[5] We assume $p_n$ is not changing with time in this case and generalize that in Appendix F.

[6] The basic reproductive number on $L_n$ must be greater than one.

[7] As another example, consider the case where the latent space is equipped with hyperbolic, rather than Euclidean, geometry (Lubold et al., 2023). While volumes in the space expand at an exponential rate, Assumption 1 may still be satisfied for some $T$ and $p_n$. In the case of sufficiently small $p_n$, this situation corresponds to the case when the diffusion simply spreads slowly because it has a low passing probability. In the case of sufficiently small $T$, this situation corresponds to the diffusion not having enough time to reach a large portion of the graph.

[8] Formally, this assumption makes sure that the diffusion does not reach the edge of particular subgraphs. Our proof strategy relies on the construction of independent subgraphs to simplify computations, so we adjust the upper bound on $T$ to compensate.

($q = 2$) is a good model of expansion, this has an upper bound of 11 months and or with a slightly more expansive model ($q = 3$), then the upper bound is 3 months. The lower bound is close to one month in either case.

We set up a general structure on the distribution of $E_n$. A given node $i$ can link to fraction $\delta_n$ of the graph through $E_n$. The share $\delta_n$ controls the support of potential links in $E_n$ and will do so in an unrestricted way. The maximum value of $\delta_n$ is clearly 1, unchanging with $n$. In this case, both "long range" and "short range" (from the perspective of $L_n$) links are permitted. But a smaller $\delta_n$ can disallow long range links – it may be that the entire $\delta_n$ share of nodes are in a highly localized neighborhood of $i$. It is useful to provide some notation to provide a lower bound on $\delta_n$. We use the same constant $a$ as in Assumption 2, with $2a > 1/(q+1)$. Let $\nu := a - 1/(2q+2)$ be the difference. The lower bound on the share of nodes that can be linked to is given by $\underline{\delta}_n = (\log n)^{-q\nu}$. Considering California again, if it were a geographic-type network ($q = 2$) or even a more expansive one ($q = 3$), it is easy to see that rates such as $\underline{\delta}_n < 0.001$ becomes permissible (as do even smaller rates). And given the unrestricted nature, this allows for topologies such as only very rare and local-in-$L$ links to be formed.

**Assumption 3.** *For every $n$, $i, j$, $E_{ij} \sim \text{Ber}(\beta_n)$ for up to some share $\delta_n$ of the $n$ nodes and is zero otherwise. Further,*

(1) $\delta_n \in (\underline{\delta}_n, 1]$
(2) $\beta_n \in \left(\frac{1}{p_n T^q \delta_n n}, \frac{1}{n}\right)$.

We can examine $\beta_n$, fixing a given value of $\delta_n$. First, note that both the upper and lower bounds for $\beta_n$ go to zero as $n$ grows large. Second, Assumptions 1 and 2 ensure that $p_n \delta_n T^q \in (1, n)$. This restriction ensures that there are links in $E_n$, with probability one, as $n \to \infty$. Third, the upper bound on $\beta_n$ imposes that $E_n$ is sparse: with probability one, $E_n$ is not a connected graph, nor will it contain a giant component as $n \to \infty$. Given these restrictions, the large forecasting errors we characterized below are not a function of a dense set of links unobserved by the econometrician. Instead, they are caused by a small (and disconnected) set of idiosyncratic links that can have an unstructured pattern. While the forecast errors would also clearly happen if the econometrician missed a dense graph or a giant component, we focus on a regime where the mismeasurement is sparse, making the results more surprising.

Econometrician's Goals. We study two policy objectives for the econometrician: identify where the diffusion goes and how much diffusion there is by time $T$.

To do so, the econometrician wants to estimate which individuals have been activated by time $T$ for a diffusion process that starts at node $i_0$ with percolation $P_n(G_n)$. Let $y_{jt}$ be an indicator which denotes if node $j$ has ever been activated through time $t$. We will generally suppress the dependence of this on $G_n$, $P_n(G_n)$, and $i_0$. The ever-activated set as of period $T$ can be written as

$$I_{P_n(G_n)}(i_0, T) := \{j \in V_n \text{ s.t. } y_{jT} = 1 | G_n, P_n(G_n), i_0\}.$$

We then consider functions of this set. We will generally assume that the econometrician is not an oracle: while they may have a (potentially large) amount of information, the realization of $P_n$ is not known. Therefore, we will study the distribution (or moments thereof) induced by the random $P_n$. As we discuss each objective function, we will make clear what is known to the econometrician. Throughout, we will assume $T$, $q$ and $L_n$ are known perfectly. These are heroic assumptions that only benefit the econometrician. Therefore, our results can be thought of as modelling policy objectives in a best-case scenario.

## 2. Sensitive Dependence on the Seed Set

We first show how the diffusion process is sensitive to weakening the assumption that the policymaker has perfect knowledge of $i_0$, the initial seed. We consider a local perturbation of the initial activation of the diffusion and show a structural lack of robustness of the model caused this initial seed mismeasurement: even if seeds differ only slightly, this is enough to generate very different diffusion patterns. The setup of the result is motivated by a policy-relevant consideration. Given slightly incorrect assessment of the seed (e.g., patient zero), even with knowledge of the true network $G_n$, the policymaker is likely to predict large differences in both *where* the diffusion jumps and *who* is activated.

To fix notation, we first define a percolation, $P_n$, for the diffusion process and vary only the initial seed between $i_0$ and some nearby $j_0$. This removes the randomness from the diffusion and holds fixed the set of possible paths that it can take as we vary the initial seed. Hence, we can study the resulting activated sets, given percolation $P := P_n(G_n)$, when seeding with some $i_0$ versus some $j_0$. Recall that $I_P(i_0, T)$ and $I_P(j_0, T)$ denote the ever-activated sets by period $T$ for the two seeds respectively, holding fixed the counterfactual passing process across each link.

We next define a catchment region. For some node $e$ that is activated at time $T$, if the diffusion process continues for $t$ more periods, then the *catchment area* is defined as the maximal set of nodes that can be activated beginning with $e$, $B_e(t)$, which is the ball centered at $e$ with radius $t$ relative to the true graph $G_n$. In what follows, we will find that, given the extreme sparsity of $E_n$, for any two nodes $e_1$ and $e_2$ which have edges in $E_n$ (i.e., there exists alters $e_1', e_2'$ that $e_1e_1',\ e_2e_2' \in E_n$), the catchment areas (over $t$ periods of transmission) typically will not intersect: $B_{e_1'}(t) \cap B_{e_2'}(t) = \emptyset$ with probability tending to one. We call $e_1'$ an alter of $e_1$ in $E_n$ as it is linked to $e_1$ in $E_n$. Intuitively, the catchment areas of these alters in $E_n$, $e_1'$ and $e_2'$, can be thought of as analogous to geographically distinct areas (though the network is not constrained to geographic structure). Each region has potential size $\mathcal{E}_t$ in expectation, and is bounded above in size by the total number of nodes in a $t$ radius ball around the seed, where $t$ is the number of periods post-seeding.

We define a sequence of *local neighborhoods relative to a diffusion process*. Let $U_{n,i_0} = B_{i_0}(a_n)$ be a ball of radius $a_n$ around the reference node, possibly growing, with $a_n/T_n \to 0$. Relative to the total expansion of the diffusion process over $T$ periods, the local neighborhood about $i_0$ we consider is vanishing.

We make use of the fact that, relative to seed $i_0$, there are two nodes, $e_1$ and $e_2$, which are the closest and second-closest nodes to $i_0$ that have a link to some respective alters in $E_n$. In what follows, we condition on the sequence of events $\Gamma_n' := \{[P_n^T]_{j_0 e_2} > 0\}$: there exists at least one path between $j_0$ and $e_2$ in the percolated graph. The construct helps us rule out pathologies and instead focus on cases where escapes are possible. In general, percolation problems with changes on linkages (e.g., bond percolation) are extremely complicated and not our focus (see, e.g., Smirnov and Werner (2001); Borgs et al. (2006)). So, we consider sequences under general conditions of interest here.[9]

We use a version of the Jaccard index (Jaccard, 1901) to compare the expected set of nodes that are ever activated by both the diffusion processes starting at $i_0$ and starting at $j_0$ relative to the number of nodes that are activated by either initial node process. We call this discrepancy measure $\Delta_n(i_0, j_0)$— the relative number of nodes that were ever activated by only one of the diffusions to the number activated by both. It is useful to also condition on the event that $i_0$ and $j_0$ are connected in the percolation, because otherwise the problem is uninteresting since the diffusions never overlap. So, we assert $\Gamma_n := \{|I_P(i_0, T) \cap I_P(j_0, T)| > 0\} \cap \Gamma_n'$ and

[9] To see an example, with infill asymptotics, one can construct sequences where $\Gamma_n$ occurs with probability tending to zero just by virtue of adding more independent paths in $L_n$ at a sufficiently high rate relative to $p_n$.

define our index as

$$\Delta_n(i_0, j_0) := \left\{ \frac{|(I_P(i_0,T) \cap I_P(j_0,T))|}{|I_P(i_0,T) \cup I_P(j_0,T)|} \,\middle|\, \Gamma_n \right\},$$

Note that $\Gamma_n$ eliminates events where the Jaccard index mechanically takes a value of zero, which makes the result that this object will be strictly less than one stronger. If $\Delta_n(i_0, j_0)$ is small for a nearby pair $i_0$ and $j_0$, then, on average with a fixed percolation, a large set of nodes is activated through the process by only one diffusion process, and not the other.

**Theorem 1.** *Let Assumptions 1, 2, and 3 hold. Let $i_0$ be an arbitrary initial seed and consider the stochastic sequence $\{G_n\}_n$ comprised of a fixed sequence of $\{L_n\}_n$ and random $\{E_n\}_n$. Let $U_{n,i_0} = B_{i_0}(a_n)$ be a ball on $G_n$ of radius $a_n$ around $i_0$ with $a_n/T_n \to 0$. Then with probability approaching one over draws of $(E_n, P_n)$, we have the following. There exists a sequence of time periods $\{T_n\}_n$, local neighborhoods $\{U_{n,i_0}\}_n$, and a sequence of shift node sets $\{J_{n,i_0}\}$ with $J_{n,i_0} \subset U_{n,i_0}$ for each $n$ with $|J_{n,i_0}|/|U_{n,i_0}| > C$ for some positive fraction $C$ independent of $n$. Furthermore, for all $j_0 \in J_{i_0}$:*

1. *The number of catchment regions disjoint from $B_{i_0}(T) \cup B_{j_0}(T)$ activated under seeding with $j_0 \in J_{i_0}$ rather than $i_0$ is order*
$$n\beta_n\delta_n p_n s_n^q > 1,$$
*for growing $s_n$, and may be order constant or even diverge in $n$.*
2. *We have*
$$\Delta_n(i_0, j_0) \leq c < 1$$
*for some fraction $c$ independent of $n$.*

All proofs are in Appendix A unless otherwise noted.

This result shows that for a non-trivial share of nodes near to $i_0$, if the seed were counterfactually shifted, we get a disjoint set of locations activated and the overall overlap is potentially low. The key idea is that when considering a diffusion pattern starting at a nearby $j_0$, we must consider whether the percolation would activate a significantly different portion of the network than the one beginning with $i_0$, by finding a "shortcut" to a very different part of the network. We show that there will always exist some $j_0$ and time period for which this is true. The intuition comes from fixing the second closest "shortcut" link in $E_n$ to $i_0$: before a diffusion pattern from $i_0$ can reach this shortcut, the diffusion from $j_0$ will reach this shortcut. This will induce two effects. First, a non-trivial share of activations will be different due to variation in seed. Figure 1 shows a heuristic construction of the set $J_{i_0}$. Second, there will be jumps in the number of distinct catchment regions activated in the network. The portion of the proof that tracks the number of catchment areas relies on a separate lemma that computes a bound on the expected number of catchment areas activated. Given this lemma will play a role in a number of results, we state it as an independent result.

**Lemma 1.** *Let Assumptions 1, 2, and 3 hold. Let $X_s$ be the number of catchment regions activated in time step $s$. Then, the following holds:*

$$\mathbb{E}_{P_n(G_n),E_n}[X_{s_n}] \geq n\beta_n\delta_n p_n s^q$$

The intuition is that the number of new catchment regions activated at each time step will be closely related to how many nodes are in the shell of diffusion on $L_n$. Note that expectations are taken over both the diffusion process $P_n(G_n)$ and realizations of $E_n$. With an application of Hoeffding's inequality, Lemma 1 yields the first part of Theorem 1.

Stepping back, we can also note that despite $j_0$ being close to $i_0$, in the sense that their network distance is small relative to the length of the diffusion, and these problems still occur. Further, these alternative seeds are not isolated: the first part of the theorem shows that a non-trivial fraction of the location neighborhood about $i_0$ contains such problematic alternative seeds. Our simulations quantify examples to show how extreme the problem can be realistic setups.

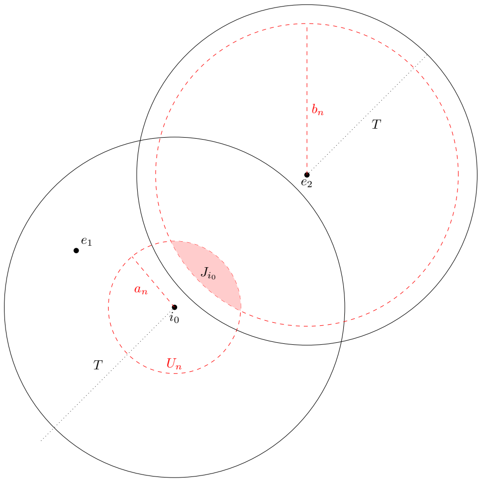


Figure 1. A heuristic construction of $J_{i_0}$ using $\mathbb{R}^2$ to represent $L_n$. Let $e_1$ and $e_2$ be the closest and second closest nodes in $L_n$ that also have a link in $E_n$. The smaller red dotted circle denotes $U_{n,i_0} := B_{i_0}(b_n)$, while the larger denotes $B_{e_2}(a_n)$. The intersection gives the set $J_{i_0}$.

## 3. Sensitivity to Mismeasurement in the Network

We now show how tiny measurement error in the network connections leads to large forecasting errors in the diffusion process. We show how using the observed network $L_n$ to make forecasts with a known seed can greatly underestimate the average extent of diffusion on the true network $G_n$.

We assume $i_0$ and $L_n$ are known perfectly. The error we study is one in which the econometrician uses the observed $L_n$ as a stand-in (mistakenly assuming $E_n \equiv 0$),

$$\hat{Y}_T(L_n) := \mathbb{E}_{P_n(L_n)}\left[\sum_{j=1}^{n} y_{jT} \;\middle|\; L_n, i_0\right]$$

where the expectation is taken with respect to the diffusion process $P_n(L_n)$ on $L_n$. We focus on this specific estimator for several reasons. First, it captures what is often done in practice both out of convenience and feasibility. Surveys about interaction and contact tracing face survey fatigue and/or top-coding in data collection. Mobility data routinely use thresholdings to define connections. They also leave out interactions that are not traceable through phones: e.g., in the developing world where households may have a single phone, movements of all members without a phone are simply dropped. In information diffusion, studies about social learning on social media may leave out person-to-person interaction, SMS off the platform, and so on. Unless one models the entire span of such missed interaction, the econometrician is really in the situation described here: effectively dropping $E_n$. Second, a consequence of some of the results below is that recovering the distribution of $E_n$ to integrate over it may be practically impossible. Even in the simple case of a homogenous $\beta_{ij,n} = \beta_n$ for every $i, j$, the imposed rates on $\beta_n$ make it difficult to identify enough links in $E_n$ to precisely estimate $\beta_n$.

Nonetheless, a reasonable if not forgiving benchmark for $\hat{Y}_T(L_n)$ is setting the target as integrating over $E_n$ rather than treating it as known.

$$\hat{Y}_T(G_n) := \mathbb{E}_{E_n, P_n(G_n)} \left[ \sum_{j=1}^{n} y_{jT} \,\middle|\, L_n, i_0 \right]$$

where the expectation is taken with respect to $P_n(G_n)$ and realizations of $E_n$, with known $i_0$ and $L_n$. If we compare the econometrician who ignores $E_n$ entirely to one who uses $E_n$ to the full extent, they will surely do worse.[10] Comparing to the case where the econometrician knows the distribution of $E_n$ and integrates over it is a more fair comparison. It also demonstrates the value of understanding the error distribution (even if it may be difficult to asses in practice).

We give the econometrician perfect knowledge of not only $L_n$, but $i_0$ as well. And, as before, the econometrician has perfect knowledge of $T$ and $q$. Despite these advantages, the econometrician's forecast error will swamp the forecast as $n \to \infty$.

**Theorem 2.** *Under Assumptions 1, 2, and 3, as* $n \to \infty$, $\frac{\hat{Y}_T(L_n)}{\hat{Y}_T(G_n)} \to 0$.

We briefly give some intuition of why the forecast error dominates the predicted error in magnitude. We count the number of catchment regions activated at each time step activated through $E_n$. While the initial missed jump to other locations in the network does not generate forecasting issues of any consequence, the issue is that these jumps recursively explode. For each catchment region activated, it may in turn activate additional catchment regions, and so on. In totality, these new shells caused by the propagating error dwarf the diffusion captured by the observed graph $L_n$.

The proof strategy formalizes this intuition. We first compute a lower bound on the number of expected new catchment regions activated in each time period with Lemma 1. To generate a lower bound on expected activations, we introduce a tiling of the graph and count how many tiles are activated in expectation. We then calculate this number and scale by the number of nodes activated in each tile.

In Appendix F, we show a similar version of Theorem 2 that allows the diffusion process to slow over time with a decays at a polynomial rate: the resulting structure is equivalent to considering a diffusion process with a lower $q$. Doing so requires slightly different assumptions: we require an earlier time window and a higher rate of missing links. The intuition is that we need additional missing links to "compensate" for the slowing diffusion process to get the same result.

## 4. Estimation and Possible Solutions

We now consider several estimation procedures in our setting. First, we consider how the econometrician can estimate the underlying structural parameters like $p_n$ successfully, despite our pathological results above. Second, we show that what seems like a natural solution to our forecasting results, estimating $\beta_n$, our error rate, and adjusting for it, is almost impossible in reasonable samples because the error rate is so small. Third, we consider a widespread testing regime and show that the detected number of regions that have activated nodes will underestimate the true number of activated regions.

[10] Furthermore, computing the expectation treating $E_n$ is known to be NP-Complete (Shapiro and Delgado-Eckert, 2012).

4.1. **Estimating Parameters of the Process.** We now show that despite the aforementioned pathologies, some core parameters of the process can be consistently estimated. We assume that the econometrician has perfect detection in that they see all true activations. Using knowledge of the observed $L_n$ and $y_{j,t-1}$, the econometrician can derive the exact number of expected activations for a given value of $p_n$ and hence consistently estimate $\hat{p}$ using the observed $y_{it}$.[11] It then follows that the econometrician will be able to consistently estimate $\mathcal{R}_0$, the basic reproduction number [12] as $\hat{\mathcal{R}}_0 = \hat{p} d_L$ where $d_L$ is the (observed) mean degree of $L_n$, whereas in actuality it is $\mathcal{R}_0(G_n) = p_n d_L + \beta_n n p_n$.

*Remark* 1. Assume that the policymaker has a consistent estimator $\hat{p}$ of $p_n$ and knows $d_L$, that $\mathcal{R}_0$ is constant, and Assumptions 1, 2, and 3 hold. Consider the estimator $\hat{\mathcal{R}}_0 = \hat{p} d_L$. Then, we have $\frac{\hat{\mathcal{R}}_0}{\mathcal{R}_0(G_n)} \to_p 1$.

*Proof.* Note that $\mathcal{R}_0(G_n) = d_L p_n + \beta_n \delta_n n p_n = d_L p_n \left(1 + \frac{\beta_n \delta_n n}{d_L}\right) = \mathcal{R}_0(L_n)(1 + o(1))$, where the final equality follows by assumption. $\hat{\mathcal{R}}_0$ is a consistent estimator of $\mathcal{R}_0(L_n)$, as $d_L$ can be computed directly and the econometrician has access to a consistent estimator of $p_n$. An application of the continuous mapping theorem completes the result. □

While the econometrician can consistently estimate $\mathcal{R}_0$ they will still be unable to accurately forecast the location or volume of diffusion as shown in Theorems 1 and 2.

We give an example of one way that an econometrician could estimate $p_n$ consistently. Let $\mathcal{I}(i,t)$ be the set of neighbors of $i$ activated at period $t$. Then at time $T$, a consistent (though inefficient) estimator of $p_n$ will be

$$\hat{p} := \sum_{t=1}^{T} \sum_{i=1}^{n} y_{it} \mathbb{1}\{y_{it-1} = 0, |\mathcal{I}(i,t-1)| = 1\} / \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbb{1}\{y_{it-1} = 0, |\mathcal{I}(i,t-1)| = 1\}.$$

Note that by restricting attention to susceptible nodes with exactly one activated neighbor, activations occur independently with probability $p_n$, and hence $\hat{p}/p_n \to_p 1$. Note that this estimator uses perfect knowledge of $L_n$ via the sets $\mathcal{I}(i,t)$.

4.2. **Possible Solutions.** We explore two possible solutions that a policymaker might pursue. First, they might estimate $\beta_n$, the connection rate for the $E_n$ graph, using supplementary measurements. Second, they might use widespread testing.

Estimating $\beta_n$. Given the previous results, an approach for the econometrician might be to estimate $\beta_n$, and use the estimate in order to inform forecasts. Assume the econometrician already has $L_n$, but is able to obtain follow-up data by sampling $m_n$ nodes uniformly at random out of $n$, and ask whether or not each $ij$ link exists in $G_n$. In this way, they can potentially find links in $E_n$ to supplement the information from the known $L_n$. Note that a sample of size $m_n$ nodes will deliver $\binom{m_n}{2}$ possible links.

We show that, in practical settings, this strategy will not be feasible. Specifically, our above theorems have demonstrated forecasting difficulties under extremely small levels of measurement error, and such a small $\beta_n$ poses challenges for estimation. Throughout, we assume that $\delta_n = 1$. We view this as a best-case scenario for the policymaker, in the sense where it makes it as easy as possible to find missing links. In fact, there are two regimes. First, with a large growing sample, the probability that one does not find a single $E_n$ link tends to one, even though the rate of $\beta_n$ is high enough to cause all the problems previously discussed. Second, one may find some missed links with a (potentially unrealistically) larger sample, but one will not be able to develop a consistent estimator.

[11] We do not solve a general formulation, as solving the generic problem is known to be NP-Hard (Shapiro and Delgado-Eckert, 2012). Rather, we show an (inefficient) estimator.

[12] The number of nodes, in expectation, activated by the first seed in an activation-free equilibrium.

**Proposition 1.** *Under Assumption 3 with $\delta_n = 1$, if:*

(1) $m_n = o(\sqrt{n})$, $\mathbb{P}\left(\text{No links amongst } \binom{m_n}{2} \text{ found}\right) \to 1$.
(2) $m_n = O(1/\sqrt{\beta_n})$, *there exists $\epsilon > 0$ and $c \in (0,1)$ such that $\mathbb{P}(|\hat{\beta}_n/\beta_n - 1| < \epsilon) < c$.*

To give a sense of scale, say that $n$ is one million. Consider a case where $\beta_n = \frac{1}{n(\log n)^2}$ (which is valid for $T = \log n$ and $q = 2$, which are allowable parameters under Assumption 2). Then, with constant $p_n$, having $m_n = o\left(\log(n) \times \sqrt{n}\right)$ samples would still deliver nearly no information, even when this corresponds to a (perfect) survey of more than 13,800 people out of $n$ equal to a million. For another example, if $\beta_n = \frac{\log n}{np_n n^{q/(1+q)}}$ (which is admissible under Assumption 2), then with constant $p_n$, under any $m_n = O\left(n \times \sqrt{\frac{1}{n^{1/(1+q)}\log n}}\right)$, an estimator for $\hat{\beta}$ is not consistent, even if there is information gained in the survey. To illustrate numerically, with the population at one million and $q = 4$, a perfect survey of nearly 68,000 people would still generate an inconsistent estimator. In practice, surveys of 15,000 people, let alone 68,000 people in a city, are uncommon. It is unlikely that this is an obstacle that can feasibly be overcome in most policy settings.

Widespread Testing. Another potential solution is the use of widespread testing. Say that a policymaker wishes to estimate where in society activated agents reside at a given time period, in order to track regions with a disease or locations susceptible to problematic rumors or where certain technologies have been adopted. We show that the number of true regions that are activated at some time period will be grossly underestimated.

Specifically, we assume that the policymaker conducts random tests instantaneously and uniformly throughout the entire society of $n$ nodes and detects the activations with i.i.d. probability $\alpha_n$. Under this widespread testing regime, we can calculate the probability that a region is correctly identified as having been seeded by period $T$ with the diffusion process.

**Theorem 3.** *Let Assumptions 1, 2, and 3 hold. Consider a test with detection probability $\alpha_n \to 0$ with $n$, such that $T < (1/\alpha_n)^{1/(q+1)}$. Let $K_T^\star$ be the expected number of regions with an activated agent at time step $T$ and let $\hat{K}_T$ be the expected number of regions with an observed activated agent at time step $T$. Assume each activated individual is observed i.i.d. with probability $\alpha_n$. Then as n $\to \infty$,*

$$\frac{\hat{K}_T}{K_T^\star} \leq \alpha_n T^{q+1} < 1.$$

This result demonstrates that in the short run, widespread testing will be bounded away from full effectiveness. The result holds because many regions will have few activations, making it harder to accurately detect them, but they will comprise a non-trivial fraction of activated regions. In practice, wide testing can become infeasible with large population. Even with an accurate test, there may not be enough tests for the full population, or it may be hard to make testing compulsory.

## 5. Extension to the Exponential Case

We turn to the case of exponential expansion, included for completeness. If there were exponential expansion, diffusion would happen so quickly that, from a policy perspective, forecasting would become moot and sensitive dependence unnecessary as the process would spread through the graph immediately. Nonetheless, we explore the implications of small mismeasurement even in this case. We make assumptions that correspond to Assumption 2 and 3, to account for the faster-moving diffusion process. As before, we assume that each node $i$ can link to a fraction of nodes $\delta_n$ of the graph through $E_n$.

**Assumption 4.** *For some constant $q > 1$ and all $t$, $\mathcal{E}_t = \Theta(q^t)$ and $\mathcal{S}_t = \mathcal{E}_t - \mathcal{E}_{t-1} = \Theta(q^t)$. In addition, we assume that $p_n\delta_n > \frac{1}{\log n}$.*

**Assumption 5.** *For every $n$, $T_n \in [\underline{T}_n, \overline{T}_n]$ where: (1) $\overline{T}_n = \log(n)$ and (2) $\underline{T}_n = \log(\log(n))$.*

**Assumption 6.** *For every $n, i, j$, $E_{ij} \sim Ber(\beta_n)$ for up to some share $\delta_n$ of nodes and is zero otherwise. Further:*

$$\beta_n = \Omega\left(\frac{1}{p_n\delta_n n}\right)$$

We then note the differences in the bounds on $T_n$: we impose a smaller lower bound and a larger upper bound than for the polynomial diffusion process. The smaller lower bound on $T_n$ is intuitive: because the diffusion spreads more quickly, the seeds from idiosyncratic links can cause the diffusion to explode much more quickly.

**Theorem 4.** *Under Assumptions 4, 6 and 5, as $n \to \infty$, $\frac{\hat{Y}_T(L_n)}{\hat{Y}_T(G_n)} \to 0$.*

We make a few comparisons to our previous result. Relative to Theorem 2, we impose a stronger lower bound on $\beta_n$ – in order for similar results to hold, we require a larger probability of idiosyncratic links. This change follows from the structure of the proof – the key comparison is the expansion in all of the areas “seeded" via the idiosyncratic links compared to the expansion of the original diffusion process. When the original diffusion process is faster moving, it means that more idiosyncratic links are needed to overwhelm the original diffusion.

Second, we note that if $p_n\delta_n < 1$, then the condition on $\beta_n$ implies that as $n \to \infty$, $E_n$ will contain a giant component almost surely. This condition will hold generically, in contrast to the case where the diffusion follows a polynomial process, which generally does not need $E_n$ to contain a giant component asymptotically. While the fraction of links missed by the policymaker still goes to zero, the policymaker still misses a large amount of structure relative to the polynomial case.

Partial Converse. With additional structure on $L_n$ we prove a partial converse to Theorem 4.

**Proposition 2.** *Assume that $L_n$ is made up of $K_n$ independent regions, which each fulfill Assumption 4. Furthermore, assume that Assumption 5 holds. Then if $\beta_n = O\left(\frac{1}{p_n\delta_n n}\right)$, we have that $\frac{\hat{Y}_T(L_n)}{\hat{Y}_T(G_n)} \to 1$.*

This result is positive for the econometrician – they correctly identify a fraction of activated nodes that asymptotically goes to 1. This follows from the fact that the initial activation creates so many activations relative to the additional “seeded” activations through $E_n$. Note that here, $E_n$ does not contain a giant component asymptotically. Combined with Theorem 4, this tells us that for a more expansive diffusion, the forecasts made by the policymaker will not be “accurate” *if and only* if $E_n$ contains a giant component. The giant component within $E_n$ will have a tree-like structure meaning that the policymaker is missing a highly expansive structure. Perfect local forecasting plays a positive role – it is what allows the policymaker to be arbitrarily accurate. But given the nature of the network structure itself, a very large share of the population becomes activated very quickly.

## 6. Simulations

We present a number of simulations to illustrate our results in finite samples and explore how variation in parameters affects things quantitatively. We simulate a Susceptible-Infected-Removed process on a network with one period of activation before removal, analogous to the processes that we study theoretically. We give an overview of each part of the simulations in the relevant subsections, with full details in the Online Appendix B.

Throughout, we fix $L_n$, the graph observed by the policymaker and design it to mimic the sparsity and clustering structure in real data. We first generate $L_n$ by placing nodes in a $q$-dimensional lattice on $[0,1]^q$. The remainder of nodes are placed uniformly at random throughout $[0,1]^q$. Nodes then link to nearby nodes, with a radius of connection chosen to ensure both that the lattice is connected and that all randomly placed nodes will be connected to the graph. As an illustrative example, we simulate two different networks with $n = 4,000$ nodes: one with $q = 4$ and one with $q = 2$. For the SIR process on the graph, we set $\mathcal{R}_0 = 2.5$, and then compute $p_n$ by dividing $\mathcal{R}_0$ by the mean degree in $L_n$. Summary statistics are shown for both graphs (along with average summary statistics for the corresponding $G_n$) in Appendix B.1.

We choose simulation time length $T$ to be twice the diameter of $L_n$ – meaning that for $q = 4$, it is chosen to be 38, while for $q = 2$ it is chosen to be 184. This value is chosen to cover both periods early on in the diffusion process, and as well as past the time period covered by our asymptotic theory.[13] Since the asymptotic theory we consider cannot speak to long-run, we simulate to the point when the diffusion extends well past the diameter of the graph, at which point we would expect the diffusion to conclude.

Sensitive Dependence. First, we investigate Theorem 1 in simulation by looking at perturbations of an initial seed within local balls covering 1% to 5% of the overall number of nodes. We fix $L_n$ and a particular instance of $E_n$ to form $G_n$, and set $i_0$ as the center of the lattice. Then, we construct $J_{i_0}$, the set of possible alternate seeds, and choose a $j_0 \in J_{i_0}$ uniformly at random. To construct $J_{i_0}$, we first find the depth of the second closest links in $E_n$ to $i_0$ – call this distance $d_{e_2}$. Then, nodes are included in $J_{i_0}$ if they are at distance $d_{e_2} + 1$ from $i_0$. Empirically, for $q = 4$, $d_{e_2} = 2$ meaning that the distance from $i_0$ to $j_0$ is 3. The local neighborhood around $i_0$, $U_{i_0}$ (which contains all nodes at or within distance $d_{e_2} + 1$) of this size makes up 5.3% of the total nodes in the graph, while $J_{i_0}$ makes up 64.6% of the local neighborhood. For $q = 2$, the distance from $i_0$ to $j_0$ is 4, while the local neighborhood of this size makes up 1.05% of the graph and the set of $j_0$ make up 31.0% of the local neighborhood.

To approximate $\Delta_n(i_0, j_0)$, we fix the underlying percolation and examine the set of ever-activated nodes infected by an epidemic that begins from $i_0$ and $j_0$. We exploit the connection between percolations and the one-period SIR process, predetermining which links in the network will transmit. However, we do not condition on the event that there is *some* overlap between the diffusions (in Theorem 1, this is encoded in the object $\Gamma_n$ and is assumed), and do not take expectations over $E_n$. However, we do take expectations over percolations. We call this version of the Jaccard index $\mathcal{J}$. We generate a single draw of $E_n$ and then hold it fixed. We simulate the process 2,500 times, and then take the average over simulations at each time period to get $\mathcal{J}(T)$.

Figures 2a and 2c indicate that there is generally little overlap between the diffusions until the process has reached the diameter of the graph and saturated the network. Recall that when $\mathcal{J}(T)$ is close to zero, this implies that the share of nodes that would be activated by both starting conditions as a share of the total activations is small. Hence, this implies that the activation paths are following very different portions of the network. This lack of overlap is despite the fact that $i_0$ and $j_0$ are extremely local. For $q = 4$, at $T = 5$ (the halfway point to the diameter of $G_n$), the value of $\mathcal{J} = 0.055$ indicates almost entirely distinct processes. For $q = 2$, at $T = 9$ (again half of the diameter of $G_n$), the value of $\mathcal{J} = 0.32$. These results are consistent with the theoretical results: there exist time periods early on in which the diffusions are almost

[13] Recall the time period bounds from Assumption 2.

entirely disjoint. Empirically, these results demonstrate that the diffusions remain disjoint for a relatively long period of time.

While it is clear that our simulations are highly sensitive to measurement error, regardless of whether $q = 2$ or $q = 4$, the changes in sensitivity are instructive. Comparing $q = 2$ to $q = 4$, the simulations demonstrate that the diffusion process is much more sensitive in terms of the extent of diffusion with lower dimension, rather than the location. This is because $q = 2$ ensures that a greater fraction of connections are "local" – therefore, there can be less local perturbation. However, i.i.d. connections lead to many more activations. Nonetheless, we note that there is still severe sensitive dependence on initial conditions with $q = 2$ – in the short run only a third of the diffusion overlaps on average.

Forecast Errors. Next, we simulate a version of Theorem 2. To do so, we simulate the error network, $E_n$, as an Erdos-Renyi graph with links that are i.i.d. with probability $\beta_n = \frac{1}{10n} = \frac{1}{40000}$, implicitly setting $\delta_n = 1$. We simulate 2,500 iterations of the SIR process on both the fixed $L_n$ and $G_n = L_n \cup E_n$, with $E_n$ re-drawn in each simulation. We do so for the $L_n$ generated with both $q = 4$ and $q = 2$. Average graph statistics for each $G_n$ are shown in Table B.1. Note that the degree distribution stays quite similar, as the average additional degree from $E_n$ is 0.100 for both sets of simulations. The initial seed $i_0$ is chosen uniformly at random and held fixed throughout the simulations. We then compute the empirical analogue of $\frac{\hat{Y}_T(L_n)}{\hat{Y}_T(G_n)}$, the ratio of the expected number of ever-activated nodes under each process.

In Figures 2b and 2d, we plot the simulated values of $\frac{\hat{Y}_T(L_n)}{\hat{Y}_T(G_n)}$ over time for each graph. For $q = 4$, the minimum ratio is attained at $T = 13$ with a value of 0.780, meaning the policymaker would underestimate the extent of the diffusion by 22%. Once the diffusion on $G_n$ reaches the diameter of the graph, the ratio increases towards a value just below one. For $q = 2$, the minimum ratio is attained at $T = 28$, taking a value of 0.169. With a lower-dimension diffusion process, the simulations are much more sensitive to additional links in $E_n$. In the Appendix B.6, we show that with $q = 2$ and $\beta_n = \frac{1}{100n}$, the minimum ratio of $\frac{\hat{Y}_T(L_n)}{\hat{Y}_T(G_n)}$ is still much smaller than the values attained with $q = 4$. The shape of the curves in Figure 2b and 2d are similar to our theoretical results, since our results focus on asymptotic results where the diffusion cannot reach the edge of the network. Hence, the ratio in our theoretical results will continue to decline. Appendix Figure B.1 shows exactly this phenomenon by separating the ratio into separate curves for $\hat{Y}_T(L_n)$ and $\hat{Y}_T(G_n)$ – the separation between the two curves is maximized just after the diameter of $G_n$ is reached.[14] Consequentially, the decline in the period prior to reaching the diameter of $G_n$ lines up exactly with the results anticipated by Theorem 2.

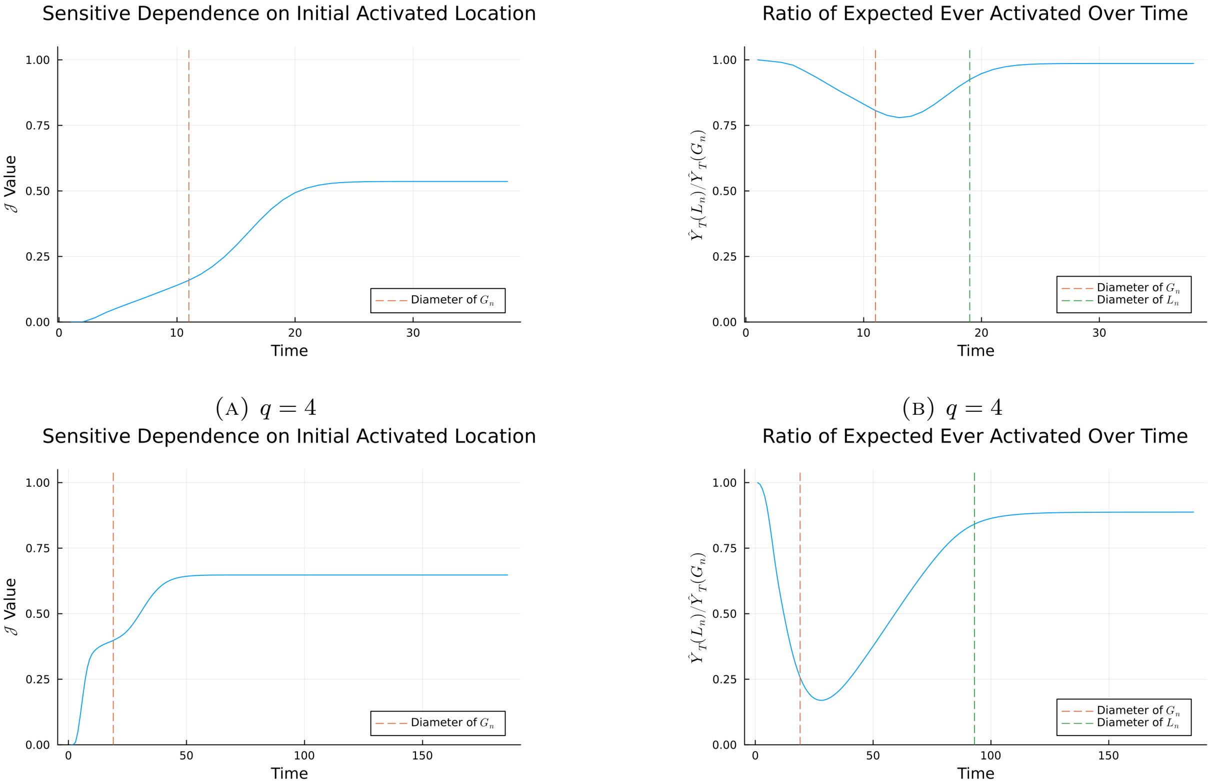
## 7. Empirical Applications

We consider three empirical applications. The first examines the COVID-19 pandemic. It demonstrates how only local linking can still cause errors in diffusion, although the problems are much worse in the idiosyncratic case. The second example studies mobile phone marketing in India, which showcases our results in a much smaller scale setting. Here, sensitive dependence on initial location has much more dramatic results – volumes of diffusion are more robust in this setting because the networks themselves are much smaller. Finally, we consider the diffusion of a weather insurance product in China. Here, we consider how errors in a diffusion model could impact statistical power when estimating peer effects.

### 7.1. **Data from the COVID-19 Pandemic.**

Kang et al. (2020) introduces a dynamic human mobility flow data set across the United States, with data starting from January 1st, 2019. By analyzing millions

[14] Note that the ratio asymptotes with $T$ to a value just below 1, as the additional links in $G_n$ allow for there to be more overall activations in expectation than in $L_n$.



(A) $q = 4$

(B) $q = 4$

(C) $q = 2$

(D) $q = 2$

FIGURE 2. Panels 2a and 2c show simulations of Theorem 1, while panels 2b and 2d show simulations of Theorem 2. Panels 2a and 2c 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 2b and 2d, 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.

of anonymous mobile phone users' movements to various places, the daily and weekly dynamic origin-to-destination population flows are computed at three geographic scales: census tract, county, and state. We study tract-to-tract flows on March 1st, 2020, at the start of the COVID-19 pandemic in the United States. Note that this date was before the WHO declared COVID-19 a pandemic and before the United States declared a national state of emergency. For the sake of computational tractability, we focus on a region of the United States that contains all of California and Nevada, along with a small portion of Arizona.

We use this real-world dataset to simulate disease transmission as in Section 6. One approach would be to construct a network with unweighted edges between two census tracts if at least one person moves between them. However, this results in an extremely dense graph. The resulting graph has a diameter of 4, a mean degree of 143.82, and a max degree of 991. The dense network will result in the epidemic spreading everywhere in a very short time, negating the need for forecasting.[15]

[15] The researcher may use the dense network and assume that $p_n$ is very small. However, with the dense network, the resulting disease process will look like an Erdos-Renyi random graph, which still follows an exponential diffusion process, rendering the forecast exercise pointless. Formally, consider the case where $G_n$ is a complete network. Then, the resulting diffusion outcome can be modeled by dropping links in $G_n$ with i.i.d. probability $1 - p_n$. The result will then be an Erdos-Renyi random graph generated with probability $p_n$, which induces exponential diffusion.

Realistically, researchers may decide to "prune" the network by only including links where there is sufficient traffic between two census tracts. In this case, a missing link implies a flow of people between two places, rather than missing a single individual contact. Hence, we construct the observed $L_n$ by linking tracts if the average flow between them (averaging over directions) is greater than six trips (the 93rd percentile of all flows). We then consider two ways to define the "true" base graph $G_n$. The first, denoted $G_n^{92}$ links tracts if the average flow exceeds five trips (the 92nd percentile), meaning that $E_n^{92}$ includes links of exactly 6 trips. Further discussion of the pruning procedure is given in Appendix C. The other, $G_n^{\beta}$, adds links i.i.d. with probability $\beta_n = \frac{1}{0.32n}$ corresponding exactly to the extra links missed going from the 5 trips to 6 trips, with these links now placed idiosyncratically. Properties of the resulting $L_n$ and $G_n$ are shown in Table C.1.

First, we simulate a version of Theorem 1, tracking $\mathcal{J}(T)$. We choose $j_0$ in a conservative fashion – after fixing a $i_0$ uniformly at random, we choose the set of potential $j_0$, $J_{i_0}$, to be all nodes at distance two from $i_0$[16]. In $G_n^{92}$, the local neighborhood containing all potential $j_0$, $U_{i_0}$, makes up 1.57% of the graph, while the set of $J_{i_0}$ makes up 81.68% of the local neighborhood. In $G_n^{\beta}$, $U_{i_0}$ contains all $j_0$ comprises 2.93% of the graph, and $J_{i_0}$ makes up 93.46% of $U_{i_0}$.

We plot $\mathcal{J}(t)$, the amount of overlap between percolations over time, in Figures 3a and 3c. These results follow the same qualitative pattern as before – $\mathcal{J}(t)$ stays close to zero for the first few time steps while the epidemics are almost entirely distinct, but then slowly increases. For the first few time periods, this graph shows dramatic sensitive dependence on the initial starting point of the epidemic. For the pruning procedure, halfway to the diameter of $G_n^{92}$, $\mathcal{J} = 0.42$. For the i.i.d. procedure, halfway to the diameter of $G_n^{\beta}$, $\mathcal{J} = 0.023$.

Next, we simulate Theorem 2 and calculate the share of $Y_t(L_n)/Y_t(G_n)$ for our two $G_n$ measures. In the first, we look at $G_n^{93} = L_n$, where $L_n$ amounts to pruning about 18 percent from the $G_n^{92}$ graph. Here, because $G_n^{92}$ is a (non-stochastic) function of the data, we hold it fixed and take expectations only over the path of the epidemic.[17] In the second, we generate $G_n^{\beta}$ via $L_n \cup E_n$, where $E_n$ has i.i.d. links to generate the same density as the error graph in the pruning procedure. In both cases, we choose $i_0$ uniformly at random and hold it fixed across simulated epidemics.

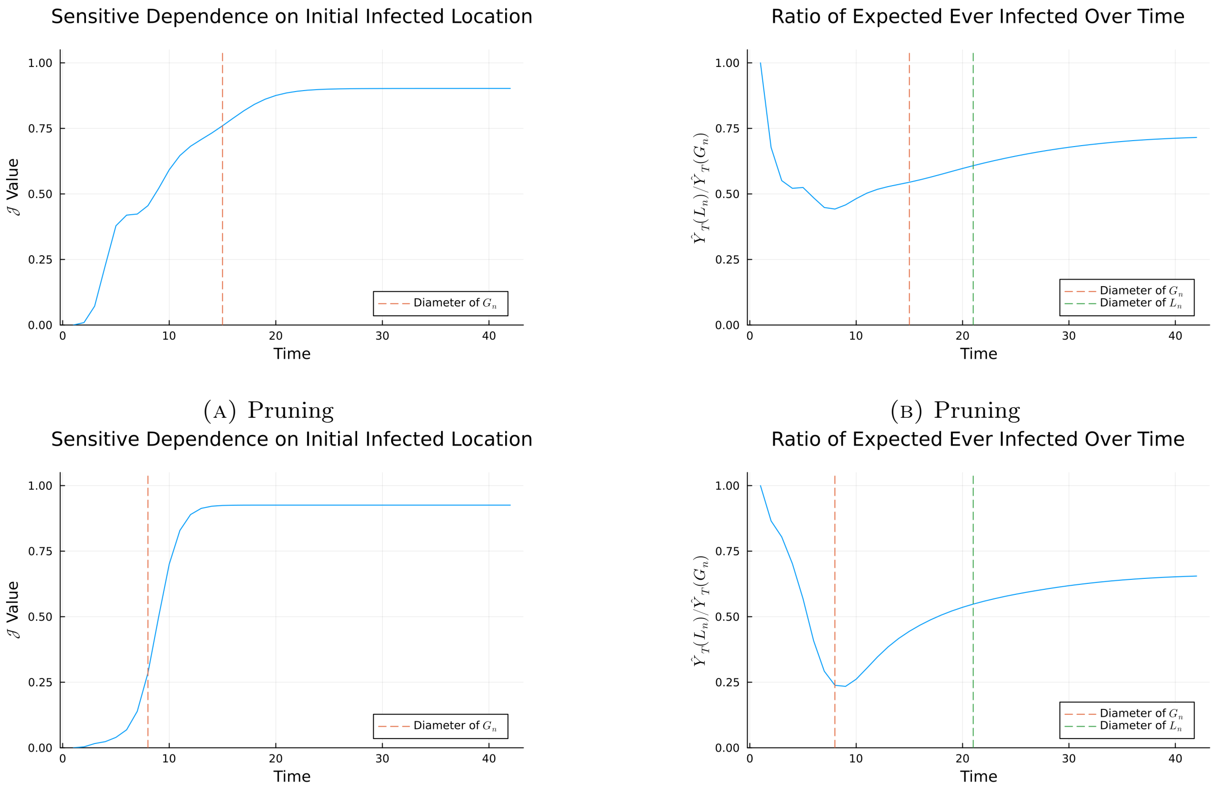
We plot $Y_t(L_n)/Y_t(G_n)$ over time in Figures 3b and 3d. For $G_n^{92}$, the pruned network, the minimum ratio of 0.442 is achieved at $T = 8$. We note that this ratio has the same qualitative pattern as in the simulated graph in Section 6 – the ratio achieves a minimum just before reaching the diameter of $G_n^{92}$, and then slowly increases. When compared to the previous simulations, the ratio increases much more slowly. This result comes from the larger dispersion in degrees – it takes longer for the disease to fully saturate the network, because there are more nodes with very few links. When compared to the i.i.d. errors in $G_n^{\beta}$, the minimum ratio of 0.234 is achieved at $T = 9$. One explanation for i.i.d. errors leading to additional underestimation follows from Theorem 2 and considering the role of $\delta_n$. The pruning procedure induces spatially clustered errors, or $\delta_n << 1$, so for the same level of error, the spatially clustered additional links in $G_n^{92}$ will not jump as far as $G_n^{\beta}$, leading to fewer "new" shells of infection.

### 7.2. **Diffusion in Mobile Phone Marketing.**

As a second empirical exercise, we study the diffusion of high-value information in Indian villages. The goal of this exercise is to highlight how the measurement issues can crop up in settings with much smaller networks, and how the initial seed condition plays a much larger role here. In Banerjee et al. (2019), one of this article's authors, along with collaborators, conducted a

[16] We found that when choosing $J_{i_0}$ based on the location of links in $E_n$, the distance from $i_0$ to the set of potential $j_0$ was typically three. Therefore, our choice of nodes at distance two is truly conservative, in the sense that we choose $j_0$ to be closer to $i_0$ than what is used in the theory.

[17] In the rest of the paper we consider expectations for Theorem 2 over both the epidemic and error graph.



(A) Pruning

(B) Pruning

(C) I.I.D.

(D) I.I.D.

FIGURE 3. Simulated version of Theorems 1 and 2 on $L_n$ and $G_n$ generated from Census tract flow data in California and Nevada. Panels (A) and (C) show simulations of Theorem 1, while Panels (B) and (D) show simulations of Theorem 2.

randomized controlled trial wherein randomly selected people in villages in Karnataka were given information on a program where they could receive a high-value cell phone or smaller cash prizes if they participated. The information about the program then diffused throughout the village.

We use this data to study the robustness of the diffusion process in an information setting. Details on how the graphs are constructed are in Appendix D. In a change from the prior simulations and analysis, many of the villages have multiple initial seeds. There are on average 3.26 seeds per village and 96 nodes per village.

We first estimate the passing probability $p_n$ for the diffusion process. Villagers could indicate they heard about the cell phone program by making a free call to the researchers. While we observe data on the sampled networks connecting households, we only observe the total number of calls received by the researchers in each village, and we do not observe whether a given household made a call. Hence, we back out the passing probability $\hat{p}_n$ using the method of simulated moments. Formally, we consider the following problem. Let $V = 69$ be the number of villages in our data (for which we have network data) and let $C_v$ be the number of calls received in village $v$. We treat the number of calls as the number of ever-activated nodes. We then simulate a SIR process with passing probability $p$ and record the number of simulated calls after $T$ periods. Let $\hat{C}_v^s(p)$ be the simulated number of calls in simulation $s$ under passing probability $p$. Then, we choose $\hat{p}_n$

as follows:

$$\hat{p}_n = \text{argmin}_p \left( \frac{1}{V} \sum_{v=1}^{69} \left( C_v - \frac{1}{s} \sum_s C_v^s(p) \right) \right) \left( \frac{1}{V} \sum_{v=1}^{69} \left( C_v - \frac{1}{s} \sum_s C_v^s(p) \right) \right)$$

We set $T = 15$, just larger than twice the average diameter of a village graph and use 2,500 simulation iterations. We estimate a value of $\hat{p}_n = 0.13$, meaning that each household transmits the information with roughly one in six chance. We then use this estimated $\hat{p}_n$ to conduct simulations.

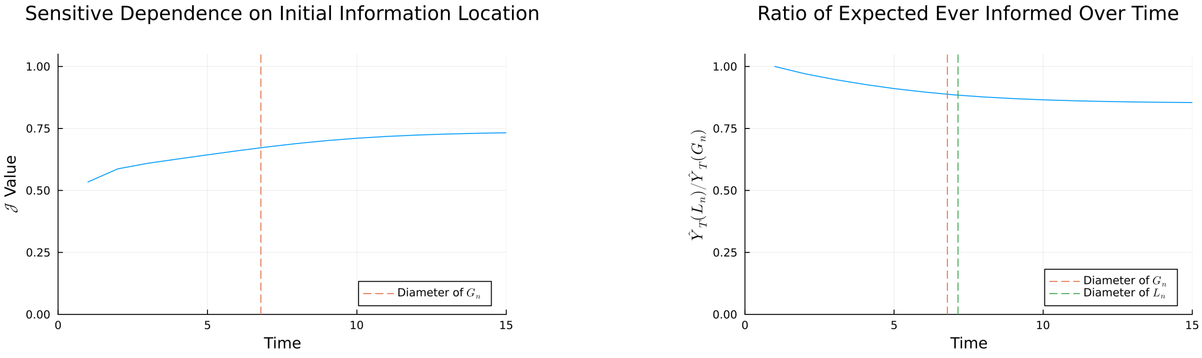
Next, we consider the error structure $E_n$ on our observed network $L_n$. Since our data has many separate villages, we consider a slightly more complex structure for $E_n$. Let $n_v$ be the number of households in village $v$. Then, we form $E_n$ by taking the union over draws of Erdos-Renyi random graphs in each village, where $\beta_n^v = \frac{1}{2n_v}$ changes in each village to keep measurement error proportional to village size. We choose a proportionally larger value of $\beta_n$ because there are multiple seeds – because the graph becomes saturated much more quickly, measurement error has less time to become a problem.

To simulate a version of Theorem 1, we choose a modified seed set for each village. Recall that most villages have multiple seeds. Here, we perturb the seed set in each village in a conservative manner. Say that a seed set is comprised of $\{i_0, j_0, k_0\}$ in some village. We choose one element of the seed set at random, say $k_0$, and then replace $k_0$ in the seed set with a neighbor chosen uniformly at random. This corresponds to a local neighborhood of 3.5% of the entire network on average. Despite the conservative perturbation, we still find similar results (Figure 4a). As before, we track $\mathcal{J}(t)$, the Jaccard index for the aggregate patterns of diffusion across all villages over time. While the value of $\mathcal{J}(t)$ does not start at 0 (as in the prior simulations), given the multiple seeds and that we conservatively only perturb one, it remains below 0.75, indicating that despite the conservative perturbation, there is still not complete overlap in the perturbed diffusion processes. Halfway to the diameter of $G_n$, the average value of $\mathcal{J} = 0.61$ indicates a lack of overlap.

We then simulate a version of Theorem 2. We simulate 2,500 diffusion processes across each village, adding up the total number of households who ever get the information and averaging across simulations. We run this both on $L_n$, the set of village graphs, and $G_n$ constructed as above (with a new draw of $G_n$ in each simulation iteration). As shown in Figure 4b, the ratio $\hat{Y}_T(L_n)/\hat{Y}_T(G_n)$ monotonically decreases over time, taking value 0.854 at $T = 15$. Despite the village-level networks being relatively small, in aggregate, the econometrician still underestimates the extent of diffusion by nearly 15 percentage points

7.3. **Treatment Effects with Spillovers in Networks.** As a third empirical exercise, we study the uptake of insurance in rural China. The goal of this exercise is to illustrate how the problems we identify in diffusion could affect conclusions from an estimated model of peer effects. If nodes are seeded with information, then the take-up behavior of a product may be a function of "exposure to information" through the diffusion process. A typical peer effects regression would consider the outcome regressed on this exposure to treatment as defined through a diffusion; our analysis suggests that results could be biased and estimators could lose considerable power.

In Cai et al. (2015), farmers receive information about a weather insurance product, a highly valuable product with low adoption rates. Intensive information sessions were randomly given to some farmers. The authors then measured the take-up by other people in the same village, who were not part of the first set of information sessions. We consider a measure of exposure to treatment based on a model of information flows.



(A)

(B)

FIGURE 4. Simulations of Theorems 1 and 2 on village networks from Karnataka, India. Panel (A) shows a version of Theorem 1. We perturb one seed uniformly at random by a single set in each village. Then, we simulate 2,500 diffusion processes on a fixed draw of $G_n$, computing the average Jaccard index of the process. Panel (B) shows a version of Theorem 2. We take 2,500 diffusion simulations on $L_n$ and $G_n$, where $G_n$ is constructed at the village level with $\beta_n = \frac{1}{2n_v}$. $n_v$ is the number of households in the village.

We first take the data from Cai et al. (2015) and build the village networks.[18] We convert the directed networks from the paper to undirected networks, where household $i$ is linked to household $j$ in our construction of the data either if $i$ reports $j$ as a link, $j$ reports $i$ as a link, or both.[19] The resulting graph is denoted $G_{n,v}$ for village $v$. Graph statistics for the villages are shown in Table E.1.

We consider an exposure measure based on a model of information flows. For a generic graph, let $A$ be the corresponding adjacency matrix. Let $s$ be a vector of indicators, with an entry equal to one if the household attended an information session. For a given $p_n$ and $T$, we define the vector of "diffusion exposure" as,

$$DE^A = \left( \sum_{t=1}^{T} (p_n A)^t \right) s,$$

which calculates the expected number of times that each individual hears information through repeated passing over $T$ periods (Banerjee et al., 2019). We imagine that the take-up of insurance in Cai et al. (2015) increases in such exposure to treatment: hearing more about the product through conversation makes one more likely to take up.[20] Note that this exposure measure is slightly different than a typical SIR model. It considers the eventual outcome as depending on the total number of times person $i$ hears about the topic through $T$ periods, rather than a once-and-for-all decision the first time someone hears about the product. This model is perhaps a more realistic description of the take-up of an insurance product. Nonetheless, the mechanics of error we outline in the paper have analogs for this kind of model.

We then simulate an experiment. We treat the data from Cai et al. (2015) as the true network $G_n$. We then regress insurance take-up ($y_{i,v}$) on the exposure measure ($DE^G_{i,v}$), a set of household controls ($X_{i,v}$),

[18] In their data collection, the authors "top-code" the number of links each household has, by only recording five outgoing links. This possibly generates measurement error as well, since it creates an artificial upper bound for all high-degree nodes, but we ignore it for our illustrative analysis (as do they in their empirical analysis).

[19] Studying an OR network may be more robust in capturing exposures due to measurement error (Banerjee et al., 2013).

[20] Following Banerjee et al. (2019), we compute this measure within each village, setting $T$ equal to the diameter of the village network. We set $p_n$ to be equal to one divided by the maximum eigenvalue of the village adjacency matrix. This is the critical value of $p_n$ such that for $p_n$ less than this value, entries of $(p_n A)^t$ tend to zero as $t \to \infty$, and some entries diverge if $p_n$ is larger.

and village fixed effects ($\mu_v$),

$$y_{i,v} = \alpha + \gamma DE^G_{i,v} + X'_{i'v}\delta + \mu_v + \epsilon_{i,v},$$

where $i$ indexes household and $v$ indexes village. To do so, we subset the data to only households who did not receive the initial informational intervention. We standardize the exposure measure to have mean zero and standard deviation one for the sake of interpretability. Results are shown in Table 1. A one standard deviation increase in diffusion exposure increases insurance uptake by 2.9 percentage points (s.e. 1.2 percentage points, $p = 0.02$), relative to a mean of 45.9%, in a linear probability model.[21]

Table 1. regression of diffusion exposure on insurance uptake

| | Insurance Uptake |
|---|---|
| Diffusion Exposure | 0.029<br>(0.012) |
| Household Controls | Yes |
| Village FE | Yes |
| Uptake Mean | 0.459 |

A regression of diffusion exposure on insurance uptake, with diffusion exposure computed from the networks collected in Cai et al. (2015). Standard errors are clustered at the village level with 2,676 observations.

We then drop links in $G_n$ with i.i.d. probability $\beta_n$ and construct $L_n$. That is, we imagine that there is a small measurement error in our survey process (or network construction process) and for this exercise we allow the error to be fully i.i.d. Our simulation corresponds to what the researcher would have observed had information flowed over $G_n$, but they instead measured $L_n$.

For each village $v$, we drop links with probability $\beta_{v,n}$, operationalized by intersecting the corresponding village graph with an Erdos-Renyi random graph with links that form with probability $1 - \beta_n$. We vary the value of $\beta_{v,n} = \frac{1}{k\bar{d}_v}$, where $\bar{d}_v$ is the village average degree and $k$ is a specified constant.[22] We vary $k$ from 5 to 15 or $\beta_{v,n}$ ranging from 0.037 to 0.0123 and recompute the diffusion exposure ($DE^L_{i,v}$), re-estimate the regression, and record the point estimate and $p$-values. We repeat this 2,500 times for each value of $k$. Let $\hat{\gamma}(G_n)$ and $\hat{\gamma}(L_n)$ be the coefficients of interest from the two regressions.

Figure 5 plots the joint distribution of the bias percentage—the percentage difference between $\hat{\gamma}(L_n)$ and $\hat{\gamma}(G_n)$—and the rejection level (one-to-one with the $p$-value) of the null of the coefficient $\hat{\gamma}(L_n)$ being equal to zero. While on average the bias is small, for any given draw we see large dispersion in the difference between $\gamma(G_n)$ and $\gamma(L_n)$ even when a very small fraction of links are dropped. In the real world, the econometrician observes only a single draw—one instance of this phenomenon. The result shows that enormous biases are possible in *any single draw*. Here, even with the smallest $\beta = 0.012$, we find the bias still has a large standard deviation of nearly 8 percentage points. With $\beta = 0.037$, biases upwards of 20% in magnitude are common.

We also see a range of p-values: as we decrease $\beta$, we would expect to see the p-values converge to the true value. With no noise, we know $p = 0.02$ and so for very small $\beta$ we might imagine that we reject the null of no peer effect at the 95% level ($0.02 < 0.05$). However, with $\beta = 0.037$, we fail to reject (at the 95%

[21] This estimated value is almost exactly half of the value reported by Cai et al. (2015) of 5.8 percentage points. Given that we use a different specification, the difference is not surprising, but it is reassuring that the results are of a similar order of magnitude.

[22] We scale $\beta_n$ by the mean degree, rather than the number of nodes, for the following reason. In order to drop a link, two things must occur: the link must exist in the first place, and that indicator must be equal to 0. In order to ensure we actually $\beta_n$ percent of links, we must scale by degree – because the graphs are sparse, if we scale by $n_v$, we drop fewer links than intended.

level) the null of no peer effects over 15% of the time. Even with $\beta = 0.012$, we still fail to reject the null of no peer effects 4.5% of the time. This means that with a very small error, between roughly 5% to 15% of the time we may be unable to reject a null at the 95% level.

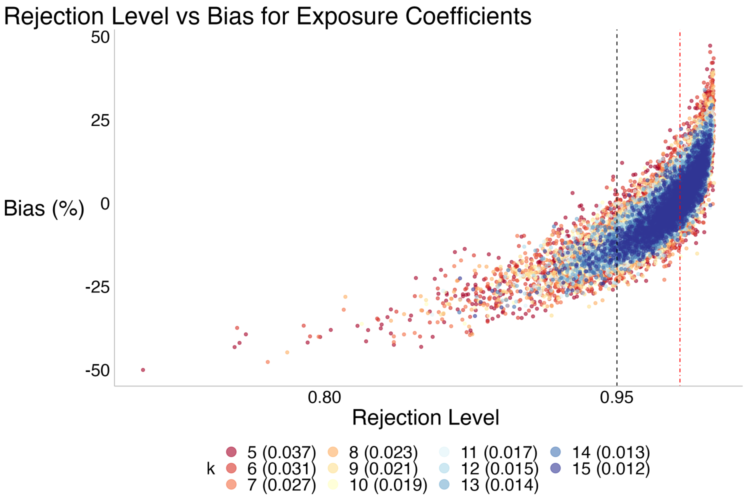


FIGURE 5. The joint distribution of the difference in $\hat{\gamma}(L_n)$ and $\hat{\gamma}(G_n)$ (in percentage terms) and the level at which we can reject the null that $\hat{\gamma}(L_n) = 0$ for different values of $k$. As $k$ increases, $\beta_{v,n}$ decreases. In parenthesis, we include the average value of the corresponding $\beta_n$ across villages. The red, dashed, vertical line denotes the level at which we can reject $\hat{\gamma}(G_n) = 0$. The black dotted line shows rejection at the 95 percent level.

## 8. DISCUSSION

We have studied the lack of robustness to extremely small quantities of mismeasurement in SIR diffusion models on networks. Such models are widely used to conceptualize epidemics, information flow, and technology adoption, among other applications. For the bulk of the paper we analyze what we call polynomial diffusion over these time horizons, capturing the idea that if it were globally exponential then the diffusion would blanket the society almost immediately. These reflect real-world contagion processes where geography, homophily, transport infrastructure, and community interactions shape the diffusion.

We have shown that even when the missed links constitute only a vanishing share of the overall links and are only concentrated locally to any node in question, the problems persist. This means that the problems are not consequences of long-range shortcuts and transitioning polynomial-like diffusion to exponential-like diffusion as in the small worlds literature. Rather, the point is that even small infrequent errors that are entirely localized wind up aggregating throughout the SIR process, thereby generating the aforementioned problems.

Our results on non-robustness are very negative about the ability to forecast diffusions, and a resourceful researcher may feel that there are solutions to circumvent these issues. For example, network data collection is known to be imperfect, but the econometrician may have knowledge of the structure of errors. In the case where the error has a known structure, the econometrician can correct for sampling error by integrating over the error (Chandrasekhar, 2016; Hsieh et al., 2024). However, this approach will only work when the error model is correctly specified; any misspecification will simply revisit the same problem. Since the problem is acute even for very small error rates, the ability to "approximately" adjust for the sampling error is insufficient. Moreover, this type of adjustment is likely missing the forest for the trees – our results give the econometrician the benefit of knowledge of a large number of typically unknown quantities and they *still* struggle to forecast diffusions as a result of this noise.

Our work demonstrates the general care needed in identifying the limits of what models can reasonably predict to inform policy. Tools must be used for exactly what they are developed. Aggregate concepts geared towards retrospective calculations may be good for just that purpose– certain aggregates, e.g., $\mathcal{R}_0$, may better be used as descriptive rather than prescriptive tools.

This raises practical concerns for any normative work that builds on the scaffolding of such models. Almost certainly the failure of robustness would propagate to welfare calculations, which often rely on the extent of diffusion or the locations (or composition or compartments) of diffusion, if not both (Acemoglu et al., 2021; Fajgelbaum et al., 2021). It is possible, though requires future work, that the susceptibility to small measurement error presents an argument for policymakers to respond earlier and much more aggressively. Barnett et al. (2023) make the point that in an uncertain world, policymakers may want to pursue more aggressive containment policies to guard against worst-case scenarios. The full decision theory exercise is beyond the scope of this paper, but it should be clear that this is the thrust of the statistical force given the massive uncertainty we document.

This paper is specific to SIR models on graphs, but the phenomenon need not be. In fact, the same sort of perturbation robustness failure may impact general models of treatment effects with spillovers (e.g., Aronow and Samii (2017), Hardy et al. (2019), and Athey et al. (2018)). The final empirical example that we presented, using the insurance take-up data from Cai et al. (2015), suggests this is exactly the case. An examination of perturbation robustness failure in general models of treatment effects with spillovers is likely worth studying in its own right which we leave to future work.

## References

Acemoglu, Daron, Victor Chernozhukov, Iván Werning, and Michael D Whinston (2021): "Optimal targeted lockdowns in a multigroup SIR model," *American Economic Review: Insights*, 3 (4), 487–502. 2, 23

Advani, Arun and Bansi Malde (2018): "Credibly identifying social effects: Accounting for network formation and measurement error," *Journal of Economic Surveys*, 32 (4), 1016–1044. 2

Alimohammadi, Yeganeh, Christian Borgs, Remco van der Hofstad, and Amin Saberi (2023): "Epidemic Forecasting on Networks: Bridging Local Samples with Global Outcomes," Tech. rep., Working paper. 4

Anderson, Roy M and Robert M May (1991): *Infectious diseases of humans: dynamics and control*, Oxford university press. 1

Aronow, Peter M and Cyrus Samii (2017): "Estimating average causal effects under general interference, with application to a social network experiment," *The Annals of Applied Statistics*, 1912–1947. 23

Athey, Susan, Dean Eckles, and Guido W Imbens (2018): "Exact p-values for network interference," *Journal of the American Statistical Association*, 113 (521), 230–240. 23

Banerjee, Abhijit, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson (2013): "The diffusion of microfinance," *Science*, 341 (6144). 20

Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson (2019): "Using gossips to spread information: Theory and evidence from two randomized controlled trials," *The Review of Economic Studies*, 86 (6), 2453–2490. 4, 17, 20, 41, 43

Barnett, Michael, Greg Buchak, and Constantine Yannelis (2023): "Epidemic responses under uncertainty," *Proceedings of the National Academy of Sciences*, 120 (2), e2208111120. 23

Borgs, Christian, Jennifer T Chayes, Remco Van der Hofstad, Gordon Slade, and Joel Spencer (2006): "Random subgraphs of finite graphs: III. the phase transition for the n-cube," *Combinatorica*, 26, 395–410. 7

Cai, Jing, Alain De Janvry, and Elisabeth Sadoulet (2015): "Social networks and the decision to insure," *American Economic Journal: Applied Economics*, 7 (2), 81–108. 4, 19, 20, 21, 23, 42, 43

Chandrasekhar, Arun (2016): "Econometrics of network formation," *The Oxford handbook of the economics of networks*, 303–357. 22

Chandrasekhar, A. and R. Lewis (2010): "Econometrics of sampled networks," MIT working paper. 2

Fajgelbaum, Pablo D, Amit Khandelwal, Wookun Kim, Cristiano Mantovani, and Edouard Schaal (2021): "Optimal lockdown in a commuting network," *American Economic Review: Insights*, 3 (4), 503–522. 2, 23

Farboodi, Maryam, Gregor Jarosch, and Robert Shimer (2021): "Internal and external effects of social distancing in a pandemic," *Journal of Economic Theory*, 196, 105293. 2

Griffith, Alan (2022): "Name your friends, but only five? the importance of censoring in peer effects estimates using social network data," *Journal of Labor Economics*, 40 (4), 779–805. 2, 3

Hardy, Morgan, Rachel M Heath, Wesley Lee, and Tyler H McCormick (2019): "Estimating spillovers using imprecisely measured networks," *arXiv preprint arXiv:1904.00136*. 23

Hoff, Peter D., Adrian E. Raftery, and Mark S. Handcock (2002): "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97:460, 1090–1098. 2, 5

Hsieh, Chih-Sheng, Yu-Chin Hsu, Stanley I.M. Ko, Jaromír Kovářík, and Trevon D. Logan (2024): "Non-representative sampled networks: Estimation of network structural properties by weighting," *Journal of Econometrics*, 240 (1), 105689. 22

Jaccard, Paul (1901): "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bull Soc Vaudoise Sci Nat*, 37, 547–579. 7

Jackson, Matthew O. (2009): "Genetic influences on social network characteristics," *Proceedings of the National Academy of Sciences*, 106 (6), 1687–1688. 1

Jackson, Matthew O and Leeat Yariv (2007): "Diffusion of behavior and equilibrium properties in network games," *American Economic Review*, 97 (2), 92–98. 3

——— (2011): "Diffusion, strategic interaction, and social structure," in *Handbook of social economics*, Elsevier, vol. 1, 645–678. 1

Kang, Yuhao, Song Gao, Yunlei Liang, Mingxiao Li, and Jake Kruse (2020): "Multiscale Dynamic Human Mobility Flow Dataset in the U.S. during the COVID-19 Epidemic," *Scientific Data*, 1–13. 15, 37

Lubold, Shane, Arun G Chandrasekhar, and Tyler H McCormick (2023): "Identifying the latent space geometry of network models through analysis of curvature," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85 (2), 240–292. 5

Newman, Mark EJ and Duncan J Watts (1999): "Scaling and percolation in the small-world network model," *Physical Review E*, 60 (6), 7332. 2

Sadler, Evan (2023): "Seeding a Simple Contagion," *SSRN paper 4032812*. 1

Shapiro, Michael and Edgar Delgado-Eckert (2012): "Finding the probability of infection in an SIR network is NP-Hard," *Mathematical Biosciences*, 240 (2), 77–84. 10, 11

Smirnov, Stanislav and Wendelin Werner (2001): "Critical exponents for two-dimensional percolation," *arXiv preprint math/0109120*. 7

Sojourner, Aaron (2013): "Identification of peer effects with missing peer data: Evidence from Project STAR," *The Economic Journal*, 123 (569), 574–605. 2

Wang, Dan J, Xiaolin Shi, Daniel A McFarland, and Jure Leskovec (2012): "Measurement error in network data: A re-classification," *Social Networks*, 34 (4), 396–409. 2

Watts, Duncan J. and Steven H. Strogatz (1998): "Collective dynamics of small-world networks," *Nature*, 393 (6684), 440–442. 2

## Appendix A. Proofs

*Proof of Lemma 1.* We can start by partitioning $L_n$ into $K$ disjoint "tiles", which generates strictly less activations than if the tiles were still connected. The tiling is a counting device – instead of counting overall activations, we count the number of tiles that are activated, and then scale those values by the number of periods for which the diffusion spreads. Each tile is composed of a subset of $L_n$ that is disjoint from every other tile.

Let $\tilde{L}_n$ be $L_n$ divided into $K$ evenly sized tiles – note that $K$ will depend on both $n$ and $T$, along with the other model primitives. We suppress this dependence for the sake of compact notation. Note that $\tilde{L}_n$ is not connected, by definition. We define $\mathcal{X}_T := \mathbb{E}_{P_n(G_n),E_n}[X_t]$, the expected number of tiles that are activated in time step $T$. We impose the following condition in the construction of the tiling for some constant $C \in [0,1)$: $C \leq \sum_{t=1}^{T-1} \mathcal{X}_t / K$ for all $T$. This ensures that there are inactive tiles for all $T$, such that we do not have saturation of the network by the diffusion. We can always construct a tiling where this is the case – by subdividing $L_n$ into balls of radius $T$ and growing $n$ sufficiently quickly relative to $T$ this will be possible. This restriction on the tiling is not entirely without loss. Instead of imposing that the diffusion does not reach the edge of $L_n$, we need to impose a bound so that it does not reach the edge of any of the tiles in $\tilde{L}_n$ – as shown in the proof, this is implied by Assumption 2.

For the sake of tractable computations, we construct a lower bound by only tracking diffusion spread in each tile that is the result of the first seed in each tile. For this simplified computation, we can compute, for $T \geq 1$:

$$\begin{aligned}
\mathcal{X}_T &= \underbrace{\beta_n \delta_n p_n}_{\text{Diffusion Jumps}} \times \underbrace{\mathcal{K}_T}_{\text{Nodes in Tiles to Jump To}} \times \underbrace{\sum_{t=0}^{T-1} \mathcal{X}_t \mathcal{S}_{T-t}}_{\text{Weight by past spread}} \\
&= \beta_n \delta_n p_n \left( n - \frac{n}{K} \sum_{t=1}^{T-1} \mathcal{X}_t \right) \sum_{t=0}^{T-1} \mathcal{X}_t \mathcal{S}_{T-t} \\
&= \beta_n p_n n \left( 1 - \frac{1}{K} \sum_{t=1}^{T-1} \mathcal{X}_t \right) \sum_{t=0}^{T-1} \mathcal{X}_t \mathcal{S}_{T-t} \\
&\approx \beta_n \delta_n p_n n \sum_{t=0}^{T-1} \mathcal{X}_t \mathcal{S}_{T-t}
\end{aligned}$$

where the approximation holds up to a constant by the construction of the tiling.

We can begin by substituting in:

$$\mathcal{X}_T = \beta_n \delta_n p_n n \sum_{t=0}^{T-1} \mathcal{X}_t (T-t)^q$$

$$= \beta_n \delta_n n p_n \left[T^q \right.$$
$$\left. + \left(\sum_{t_1=1}^{T-1} (T-t_1)^q \left(\beta_n \delta_n p_n n \left[t_1^q + \sum_{t_2=1}^{t_1-1} (t_1-t_2)^q \left(\beta_n \delta_n p_n n \left[t_2^q + \times ...\right]\right)\right]\right)\right)\right]$$

Note that the nested summation must be polynomial in $T$, despite the multiplicative structure. While we have combinatorial growth in the number of terms, we are only multiplying polynomials of $T$ together. As polynomials are closed under multiplication, the result will be a polynomial in $T$, with the lead term to be $T^q \beta_n \delta_n n p_n$.

To complete the proof, we verify the validity of the tiling with the given assumptions. We verify compatibility with Assumption 2. First, note that to have links in $E_n$, in expectation, we must have:

$$p_n \delta_n T^q < n \Rightarrow T < \left(\frac{n}{p_n \delta_n}\right)^{1/q}$$

Second, recall the assumption we made in the tiling: we have to be able to divide $L_n$, the base graph, into enough tiles. We can collect the relevant conditions:

$$K(T,n) \geq \sum_{t=0}^{T-1} \mathcal{X}_t \geq \mathcal{X}_{T-1} \geq \beta_n \delta_n p_n n (T-1)^q,$$
$$n > K(T,n)\mathcal{E}_T \Rightarrow \frac{n}{T^{q+1}} > K(T,n)$$

The first statement holds by construction and evaluating based on prior computations. The second statement enforces that the total expected number of activations in all tiles must be less than $n$ – mechanically, this enforces that not all nodes are activated in expectation. We can combine inequalities to get $\frac{n}{T^{q+1}} > \beta_n \delta_n p_n n (T-1)^q$. Given that $\beta_n > \frac{1}{p_n \delta_n n T^q}$, asymptotically this gives us that $T < n^{\frac{1}{q+1}}$. This is the stricter of the two upper bounds on $T$, so it binds (and is exactly the upper bound of Assumption 2).

We can consider the resulting structure of the tile level graph, despite $E_n$ not necessarily being connected. This will give us a lower bound on $T$, as we implicitly assume that the tile level graph to be connected with probability one. We imposed that there are $v(T,n) = n/K(T,n)$ nodes per tile. Given $\beta_n$, the probability of connection between two *tiles* will be $1 - (1-\beta_n\delta)^{v(T,n)^2} \approx \beta_n \delta v^2(T,n)$. We want this quantity to be at least $\log n/n$. Re-writing our expression for the tile link rate in terms of $K$ yields the following expression.

$$\beta_n \delta_n \frac{n^2}{K(T,n)^2} > \frac{\log n}{n} \Rightarrow \beta_n \delta_n > \frac{\log n}{n^3} K(T,n)^2.$$

We can then consider this expression when $\beta_n$ is as small as possible, and $K(T,n)$ is as large as possible, and note that this is consistent with Assumption 2 that

$$\frac{1}{p_n n T^q} > \frac{\log n}{n} \frac{1}{T^{2q+2}} \implies T > (p_n \log n)^{1/(q+2)}$$

Note that this is a (much) stricter lower bound that what is imposed by Assumption 2. Thus the tighter lower bound will still give the desired properties. This completes the proof. □

*Proof of Theorem 1.* Fix the percolation $P_n$ and recall in what follows $\Gamma_n$ is respected. All distances are with respect to $P_n \cap G_n$, meaning the intersection of the realized graph and the realized percolation. Recall that $e_1$ is the closest node to $i_0$ in $P_n$ that also has a link in $E_n$. Let $e_2$ be the second closest such node.

Define $r := d(i_0, e_2)$, the distance between $i_0$ and $e_2$. Set $T = \kappa \cdot r$ for some $\kappa > 0$, which determines the diffusion duration. Then let $a_n = o_p(r)$ growing in $n$ be a distance and $U_n := B_{i_0}(a_n)$. Note $|U_n|/T_n^{q+1} \rightarrow_p 0$ by construction, meaning that $U_n$ is a sequence of local neighborhoods vanishing relative to the diffusion.

Then pick $b_n = r - ca_n$ for $c \in (0,1)$, constant in $n$. Notice the lens formed, $\ell(a_n, b_n; r) := U_n \cap B_{e_2}(b_n)$ is of constant order relative to $U_n$. Let $J_{i_0} := \ell(a_n, b_n; r)$, completing the construction of $J_{i_0}$.

We can then prove the first part of the Theorem. Every $j_0 \in J_{i_0}$ reaches $e_2$ with at least $s_n = cb_n - 1$ more steps. At that point at least $s_n^q$ activations occur about alter $e_2'$ of $e_2$. We can think of a new diffusion starting at $e_2$ for at least $s_n$ periods. The region around the alter of $e_2$ will be the first region seeded, and there will be potentially more in expectation, depending on the parameters. By Lemma 1, the number of regions activated in expectation will be at least: $n\beta_n\delta_n p_n s_n^q$. Recall that this result relies on choosing a tiling with $K$ regions – we take the regions to be the catchment areas themselves. Note that $K$ is growing in $n$. Then, it follows that:

$$\mathbb{P}(|X_{s_n} - n\beta_n\delta_n p_n s_n^q| \geq K\epsilon) \leq 2\exp\left(-2\epsilon^2 K\right) \to 0$$

Via an application of Hoeffding's inequality to the set of indicators for if a catchment region has been activated. This completes the first part of the proof.

We can show that $\Delta_n(i_0, j_0) < c < 1$ for some positive fraction independent of $n$. For any $P$, the distance between the two nodes is order $b_n$, so the lens between them has order $b_n^q$ as does the disjoint set. But this is the same order as $s_n^q$ which we saw as the volume of the activations emanating from alter $e_2'$. So the result follows as this holds for any $P$ that respects $\Gamma_n$. □

*Proof of Theorem 2.* We can first note that the numerator is exactly $\hat{Y}_T(L_n) = \mathcal{E}_T$ and can be bounded from above using Assumption 1. Then, we can construct a tiling and apply Lemma 1.

Formally:

$$\hat{Y}_T(G_n) = \mathbb{E}\left[\sum_{j=1}^{n} y_{jT} = 1 \;\middle|\; E_n + L_n\right] \geq \mathbb{E}\left[\sum_{j=1}^{n} y_{jT} = 1 \;\middle|\; E_n + \tilde{L}_n\right].$$

The lower bound comes from ignoring the spread between tiles – instead, we only allow for inter-tile spread through $E_n$. We will lower bound the expression further by only counting the first activation in each tile.

Note that Lemma 1 provides a lower bound for the number of *tiles* seeded in each period (only tracking first activations), but we want the number of nodes ever activated. This will be $\sum_{s=0}^{T} \mathcal{X}_s \mathcal{E}_{T-s}$, where we weight the spread in each period $\mathcal{E}_{T-s}$ by the number of tiles seeded for the first time in that period. We must weight the number of tiles by the volume of (expected) spread given the initial activation time. Therefore we have the following:

$$\begin{aligned}
\hat{Y}_T(G_n) &= T^{q+1} + \sum_{s=0}^{T-1} \mathcal{X}_s (T-s)^{q+1} \\
&\geq T^{q+1} + \beta_n p_n n \sum_{s=0}^{T-1} s^q (T-s)^{q+1} \\
&\geq T^{q+1} + \frac{1}{4^{2q+1}} \beta_n p_n n T^{2q+1}
\end{aligned}$$

where the second bound comes from taking only the term corresponding to $\frac{T}{2}$ from the sum, which will be the largest individual term.[23]

[23] We assume for the sake of more compact notation that $T$ is even – if odd, simply take the floor of $T/2$ and the order of magnitude and thus the proof is preserved.

Now we can consider our object of interest using these bounds:

$$\frac{\hat{Y}_T(L_n)}{\hat{Y}_T(G_n)} \leq \frac{T^{q+1}}{T^{q+1} + T^{2q+1}\beta_n\delta_n p_n n/4^{2q+1}} = \frac{1}{1 + T^q\beta_n\delta_n p_n n/4^{2q+1}}$$

Then, by Assumption 3, this quantity will go to 0 as $n \to \infty$ and $T \to \infty$. □

*Proof of Proposition 1.* For (1), We note as $m_n = o(\sqrt{n})$ and $\beta_n \in \left(\frac{1}{p_n n T^q}, \frac{1}{n}\right)$, then $\beta_n m_n = o\left(\frac{1}{\sqrt{n}}\right)$, $\beta_n m_n^2 = o(1)$. Then we have that

$$\begin{aligned}\mathbb{P}\left(\text{No links amongst } \binom{m_n}{2} \text{ found}\right) &= (1-\beta_n)^{\binom{m_n}{2}} \approx 1 - \beta_n \binom{m_n}{2} \\ &= 1 - \beta_n \frac{m_n^2 - m_n}{2} = 1 - o(1) + o\left(n^{-1/2}\right) \to 1,\end{aligned}$$

where we use the binomial approximation. Note that this will tend to 1 even in the most adversarial case, where $\beta_n$ is as large as possible ($m_n = o(\sqrt{n})$).

For (2), it suffices to show that a necessary condition for the law of large numbers fails. Let $e_{ij}^n$ denote a potential edge in $E_n$ and $z_{ij}^n = e_{ij}^n/\beta_n$ which is a normalized version. Then we can calculate, for $s_{ij}$ a dummy for the pair being sampled,

$$\operatorname{var}\left(\frac{2}{m_n(m_n-1)} \sum_{i,j:s_{ij}=1} z_{ij}^n\right) = \frac{1}{\beta_n^2}\frac{2}{m_n(m_n-1)}\beta_n(1-\beta_n) = \frac{2(1-\beta)}{m_n^2\beta_n - m_n\beta_n}.$$

For the law of large numbers to apply we need the variance to go to zero and therefore we need $m_n^2\beta_n$ to diverge, and this fails under the hypothesized condition. □

*Proof of Theorem 3.* We assume the policymaker observes an activated agent with a known probability $\alpha_n$. The total number of activations can be accurately estimated by dividing the observed total count by $\alpha_n$. Say that a region has $x$ activations: then the probability of at least one activation being detected will be $1 - (1-\alpha_n)^x \approx \alpha_n x$. Because this expression is approximately linear, the probability of detecting at least one activation in period $t$ will be $\Theta(\alpha_n t^{q+1})$ via Assumption 1. We then want to scale by the number of regions activated in each period. This is exactly analogous to Lemma 1. Here, we take the tiles used in the proof to be the regions themselves. Recall that at time $T$ there will be at least $\beta_n\delta_n n p_n T^q$ regions activated in expectation – lower bounding $K_T^*$. So we have that

$$\frac{\hat{K}_T}{K_T^*} \leq \frac{\alpha_n\beta_n\delta_n n p_n T^{2q+1}}{\beta_n\delta_n n p_n T^q} + \alpha_n \frac{o\left(T^q\right)}{\beta_n\delta_n n p_n T^q} \leq \alpha_n T^{q+1} < 1.$$

as $n \to \infty$, which completes the proof. □

*Proof of Theorem 4.* We can begin with a similar computation to the polynomial case, though the exponential nature of $\mathcal{E}_t$ makes exact computations possible. We begin with the analogue of Lemma 1, again working with a tiling of $L_n$. Again assuming that $K(T)$, the number of tiles, grows sufficiently quickly we can compute:

$$\mathcal{X}_T \geq \beta_n p_n n \sum_{t=0}^{T-1} \mathcal{X}_t \mathcal{S}_{T-t} = \beta_n\delta_n p_n n \sum_{t=0}^{T-1} \mathcal{X}_t q^{T-t} = \beta_n p_n \delta_n n (1 + \beta_n\delta_n p_n n)^{T-1} q^T$$

Then, we can compute:

$$\frac{\hat{Y}_T(L_n)}{\hat{Y}_T(G_n)} \leq q^T \left(q^T + \sum_{s=0}^{T-1} \beta_n\delta_n p_n n (1 + \beta_n p_n n)^{s-1} q^T q^{T-s}\right)^{-1}$$

$$= \left(1 + \frac{(1+\beta_n\delta_n np_n)^T - 1}{1+\beta_n\delta_n np_n}\right)^{-1}$$

This quantity then goes to zero by Assumption 6.

We can then verify the validity of the tiling. We begin with our conditions on the tiling and that not all nodes are activated in expectation.

$$K(T,n) \geq \sum_{t=0}^{T-1} \mathcal{X}_t \geq \mathcal{X}_{T-1} = \beta_n\delta_n p_n n(1+\beta_n\delta_n p_n n)^{T-2}q^{T-1}, \text{ and } \frac{n}{q^T} > K(T,n)$$

in an identical fashion to the proof of Theorem 2. We can chain inequalities to get:

$$\frac{n}{q^T} > \beta_n\delta_n p_n n(1+\beta_n\delta_n p_n n)^{T-2}q^{T-1}$$
$$\log(n) > \log(\beta_n\delta_n p_n n) + (T-2)\log(1+\beta_n\delta_n p_n n) + (2T-1)\log(q)$$

By Assumption 6, we have that $\beta_n\delta_n p_n n > \varepsilon > 0$ so the bound reduces to $T = O(\log n)$. This restriction is exactly the first part of Assumption 5. For the second part of the bound, we repeat the same computation from the proof of Theorem 2, ensuring that the tile level graph is connected almost surely. We know that the following must hold:

$$\beta_n\delta_n > \frac{\log n}{n^3}K(T,n)^2 \implies \frac{1}{p_n n} > \frac{\log n}{n}\frac{1}{q^{2T}} \implies q^{2T} > p_n \log n$$
$$2T\log(q) > \log(p_n) + \log\log(n) \implies T > \frac{\log p_n}{2\log q} + \frac{\log\log n}{2\log(q)}$$

So the key condition is $T = \Omega(\log\log(n))$, which is exactly the second condition on $T$ from Assumption 5. Note that we use Assumption 4 so that this bound is well-defined. This completes the proof of the Theorem. □

*Proof of Proposition 2.* Recall that under Assumption 4, and $L_n$ being divided into $K(T,N)$ independent tiles, we can compute the expected number of regions activated at time $T$ via a recursion in the same way as before: $\mathcal{X}_T = \beta_n p_n n(1+\beta_n p_n n)^{T-1}q^T$. Note that because we assume $L_n$ is divided into tiles, the computation is exact rather than a lower bound. Note that tracking secondary activations preserves the same order of magnitude. Then, by the same computation as before we have:

$$\frac{\hat{Y}_T(L_n)}{\hat{Y}_T(G_n)} \geq \left(1 + \frac{(1+\beta_n\delta_n np_n)^T - 1}{1+\beta_n n\delta_n p_n}\right)^{-1}$$

which goes to 1 as $\beta_n = O\left(\frac{1}{p_n n}\right)$. Verification of the tiling strategy proceeds in much the same way as in the proof of Theorem 4, which completes the proof. □

# Online Appendix

## Appendix B. Simulation Details

To illustrate and expand on the results from the main text, we run a number of simulations. Here, we describe the simulations in detail.

### B.1. 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 1. 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 $[0,1]^q$.

(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 $\beta_n$. 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 B.1 in the main text, and for additional simulations in Table B.2.

Table B.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 | 11.609 | 93.0 | 20.439 |
| Mean Degree | 10.164 | 10.263 | 5.826 | 5.926 |
| Min Degree | 3.0 | 3.095 | 2.0 | 2.0 |
| Max Degree | 24.0 | 24.103 | 16.0 | 16.13 |
| Mean Clustering Coefficient | 0.265 | 0.258 | 0.379 | 0.37 |
| Average Path Length | 7.548 | 6.018 | 31.807 | 10.312 |

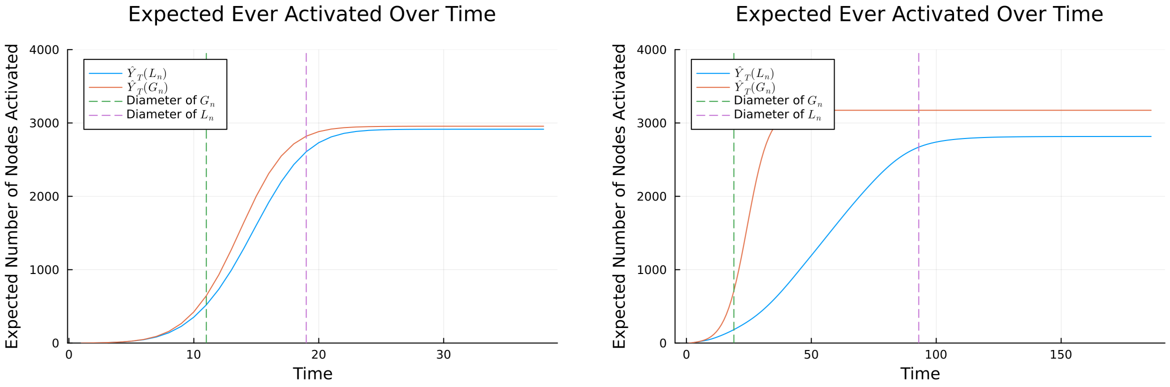
For $q = 4$, 60 percent of nodes are in the lattice, while with $q = 2$ 62.5 percent are. Statistics for $G_n$ are the expectation over 2,500 draws of $E_n$, which is drawn Erdos-Renyi with $n = 4,000$ and $\beta_n = \frac{1}{10n} = \frac{1}{40000}$.

### B.2. 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 $L_n$.

### B.3. 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 $J_{i_0}$. First, we find the second distance of the closest link in $E_n$ – denote this $d(e_2)$. Then, all nodes at $d(e_2) + 1$ are included in $J_{i_0}$. We then choose a $j_0 \in J_{i_0}$ uniformly at random.



(A) $q = 4$

(B) $q = 2$

FIGURE B.1. This figure plots the same information as Figure 2, 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 2. As $\hat{Y}_T(L_n)$ catches up, the ratio increases.

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-infected nodes under the epidemic from $i_0$, and $I_P(j_0)$ be the corresponding set from $j_0$. Then, we compute:

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

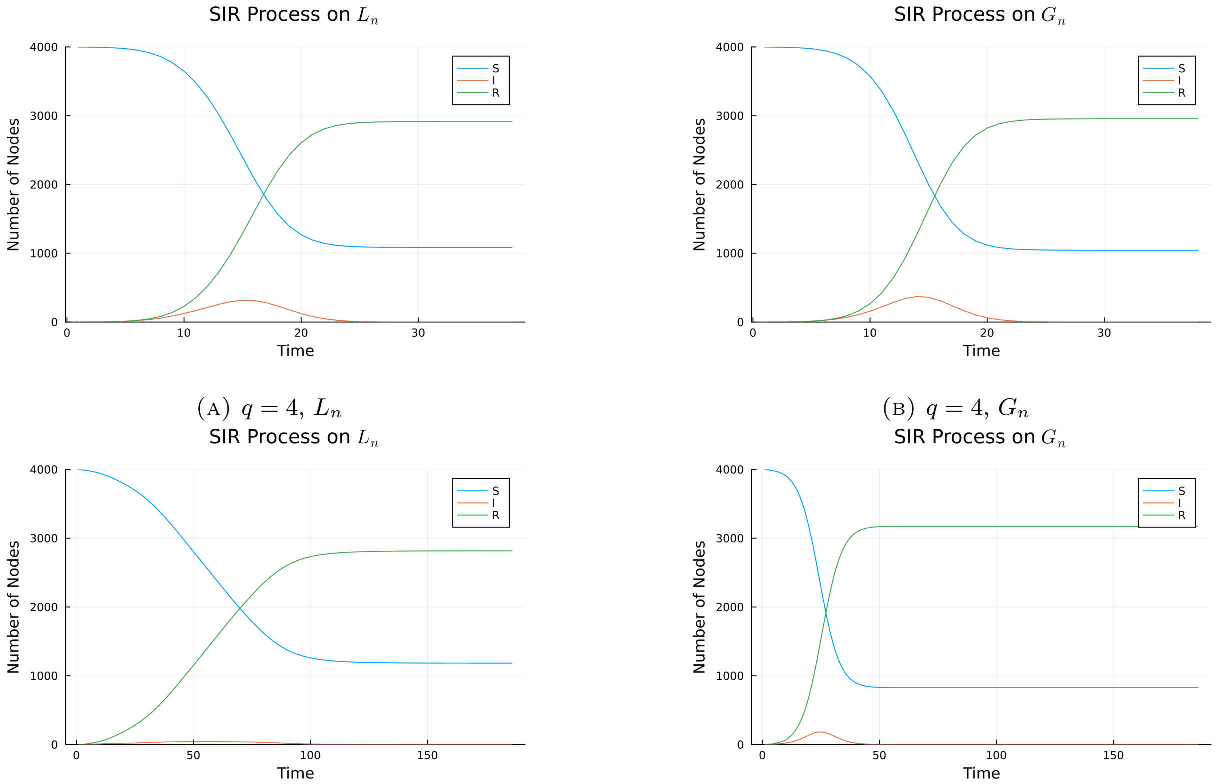
We define the Jaccard index $\mathcal{J}$ in a slightly different fashion than to $\Delta_n$, the Jaccard index in Theorem 1. We no longer condition on $\Gamma_n$, which ensures some overlap. In addition, we consider the expectation over draws of the diffusion process.

B.4. **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 $\beta_n$.
(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. In the main text, results are shown in Figure 2. Additional results are shown in Figures B.1 and B.2.

B.5. **Aggregate Patterns Are Well-Approximated by Compartmental Models.** As an additional exercise, we study the approximation of the diffusion process by a standard differential equations SIR compartmental model. Instead of the network based SIR model, we assume the policymaker estimates



(A) $q = 4$, $L_n$

(B) $q = 4$, $G_n$

(C) $q = 2$, $L_n$

(D) $q = 2$, $G_n$

Figure B.2. 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.

the parameters of a version of the standard differential equation SIR model. Changes in the number of susceptible ($S(t)$), infected ($I(t)$), and removed ($R(t)$) at time $t$ are given by:

$$\begin{aligned}\dot{S}(t) &:= -\frac{s}{n}S(t-1)I(t-1)\\ \dot{I}(t) &:= \frac{s}{n}S(t-1)I(t-1) - rI(t-1)\\ \dot{R}(t) &:= rI(t-1)\end{aligned}$$

Where $s$ and $r$ are parameters that govern the disease process. Note that $\mathcal{R}_0 = s/r$. This model is exactly a discrete-time analogue of the standard SIR model.

We assume that the policymaker estimates $\hat{s}$ and $\hat{r}$ from observed data via a set of moment conditions, matching both the number of infected and removed people at each time step. It will be useful to define some additional notation. Let $N$ be the number of simulations. Let $I_n^s(t)$ be the number of infected people at time $t$ in simulation $n$. Let $R_n^s(t)$ be defined analogously for recovered. Let $I(t; s, r)$ be the number of infected at time $t$ with parameters $r$ and $s$. Let $R(t; s, r)$ be defined analogously. Then, the policymaker solves the following problem for each simulation run, given $T$ periods of data. We then collect the moment conditions in the following vector:

$$M_n(t) = \begin{pmatrix} I_n^s(t) - I(t;s,r) \\ R_n^s(t) - R(t;s,r) \end{pmatrix}$$

Then the policymaker solves:

$$\{\hat{s}_n, \hat{r}_n\} := \operatorname{argmin}_{s,r} \frac{1}{T} \sum_{t=1}^{T} M_n(t)' M_n(t)$$

For each simulation. Then, we compute the following quantities, getting the average trajectory from the fitted SIR models.

$$\bar{I}(t) = \frac{1}{N} \sum_{n=1}^{N} I(t; s_n, r_n), \quad \bar{R}(t) = \frac{1}{N} \sum_{n=1}^{N} R(t; s_n, r_n), \quad \mathcal{R}_0 = \frac{1}{N} \sum_{n=1}^{N} \frac{s_n}{r_n}$$

We can also compare directly to the metric of average ever activated, our policy object of interest for much of the main text, by computing $\bar{I}(t) + \bar{R}(t)$ at each time period.

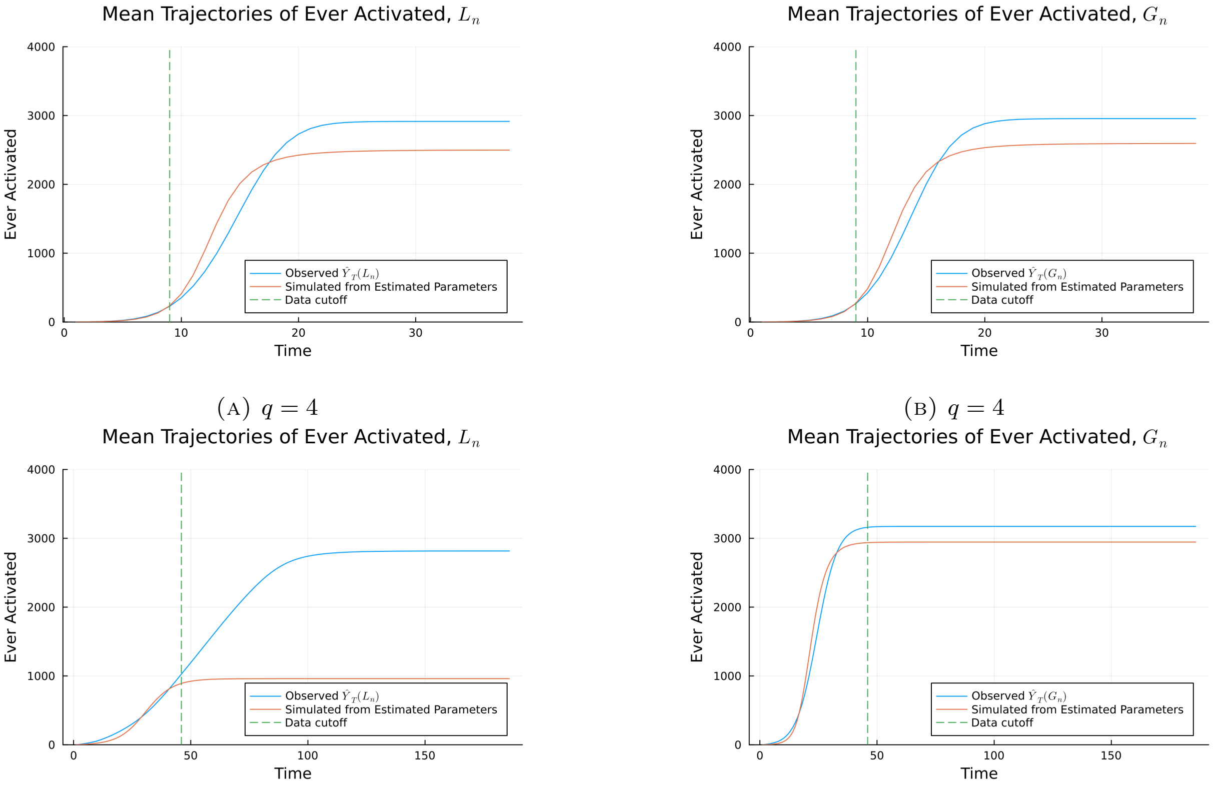
We conduct two exercises. In the first exercise, we simulate a diffusion process on $G_n$ for $T$ periods. We then estimate the parameters of interest, $(\hat{r}, \hat{s})$ at $\hat{t} = T/4$ and we generate forecasts from the compartmental model. We compare this to the actual diffusion trajectory. The second exercise replicates the first, with the only change being that we simulate the diffusion process on $L_n$ instead. Note that this is not what generates the diffusion process in the "real world"—that is diffusion on $G_n$. However, together the two simulations capture two features: (a) the deviation of the mean-field model from the underlying discrete process and (b) how the deviation depends on the relative structure of $G_n$ to $L_n = G_n - E_n$. We repeat both sets of simulations for both $q = 4$ and $q = 2$.

Figure B.3 presents the results. We begin with $q = 4$ and it is helpful to look to the diffusion on $L_n$ first in Panel B.3a. This shows how well the mean-field approach captures the dynamics on a network structure ignoring links in $E_n$. In the periods where the SIR process is fit to the simulated data, the fit is very good. The estimated $\hat{\mathcal{R}}_0$, derived by taking the average across simulations of $\hat{s}/\hat{r}$, is 1.46 under $\hat{Y}_T(L_n)$, well below the true $\mathcal{R}_0$ of 2.5.[24] The estimated forecasts (in orange) diverge quickly from the true diffusion, $\hat{Y}(L_n)$. Because of the initially exponential growth structure of the compartmental model, early in the medium run it overshoots, though the diffusion saturates much earlier and in fact the overall diffusion count in the long run is underestimated.

That is, in sample, the compartmental model can be made to fit well, but with a lower growth rate for the number of ever-infected nodes. However, because of the lower implied $\mathcal{R}_0$, the compartmental model dramatically underestimates the total number of expected activations out of sample. Ex-post, a policymaker could fit this type of model and do extremely well, but it would not be helpful for predicting the future trajectory.

In Panel B.3b, we turn to diffusion on $G_n$. The estimated $\hat{\mathcal{R}}_0$, derived by taking the average across simulations of $\hat{s}/\hat{r}$, is 1.52 under $\hat{Y}_T(G_n)$, still below the true $\mathcal{R}_0$ of 2.5. We find very similar results as the case with $L_n$. The principal difference is that the idiosyncratic links, $E_n$, generate a slightly closer forecast curve to the true trajectory. While the in-sample estimates are quite good, the exponential structure makes the process run too fast and then fade too early as well, relative to a slower more persistent polynomial process.

[24] Note that while Lemma 1 implies that there exists a consistent estimator of $\mathcal{R}_0$, the estimator we propose in theory uses activation-level data. Here, we base our estimate of $\hat{\mathcal{R}}_0$ using the aggregate diffusion pattern.



(A) $q = 4$

(B) $q = 4$

(C) $q = 2$

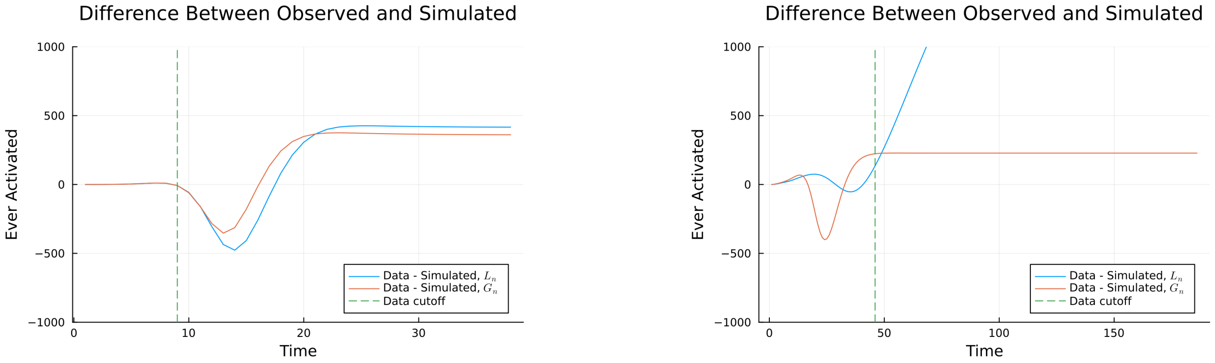
(D) $q = 2$

FIGURE B.3. A comparison of the mean ever activated under the true network SIR model and the estimated trajectory from the differential equations model. Panel (A) and (B) use $q = 4$, while (C) and (D) use $q = 2$. Panel (A) shows simulations when $\hat{Y}_T(L_n)$ is used as the data generating process, while Panel (B) shows when $\hat{Y}_T(G_n)$ is used. The data cutoff is at $T/4$. Before this point, the compartmental SIR model is fit to the data via the generalized method of moments. After this point, we run the differential equation model forward out of sample.

When we reduce to $q = 2$, there is a shift between $L_n$ and $G_n$, as seen in Panel B.3c. Now, the process cannot be well approximated by the model. The fitted compartmental SIR looks almost nothing like the true trajectory: while fitting to data the SIR model makes a complete "S" curve shape, it dramatically underestimates the total activations. In Panel B.3d, the compartmental SIR model is able to match the data more closely, because the diffusion moves much more quickly.

For $\hat{Y}_T(L_n)$, the average (across simulations) root mean squared error (RMSE) is 11.43, while with $\hat{Y}_T(G_n)$ it is 11.89. Unsurprisingly, the RMSE under $\hat{Y}_T(G_n)$ is larger, as the data is inherently noisier. The simulated trajectories quickly diverge from the data out of sample. In the next $T/4$ periods, the average RMSE with $\hat{Y}_T(L_n)$ is 429.08, while with $\hat{Y}_T(G_n)$ it is 354.21. This divergence is shown in Figure B.3.

As an additional exercise, we plot the difference between the simulated forward and "true" trajectories under each data generating process. Results are shown in Figure B.4. We can note that under the true data generating process of $\hat{Y}_T(G_n)$, the maximum under and over-estimation by the SIR differential equation model is smaller than under $\hat{Y}_T(L_n)$. The additional i.i.d. links increase the degree of the polynomial, meaning that an exponential SIR model can more closely approximate the process. This effect is much larger with $q = 2$ than with $q = 4$, as this is when the SIR model approximates the process more poorly.

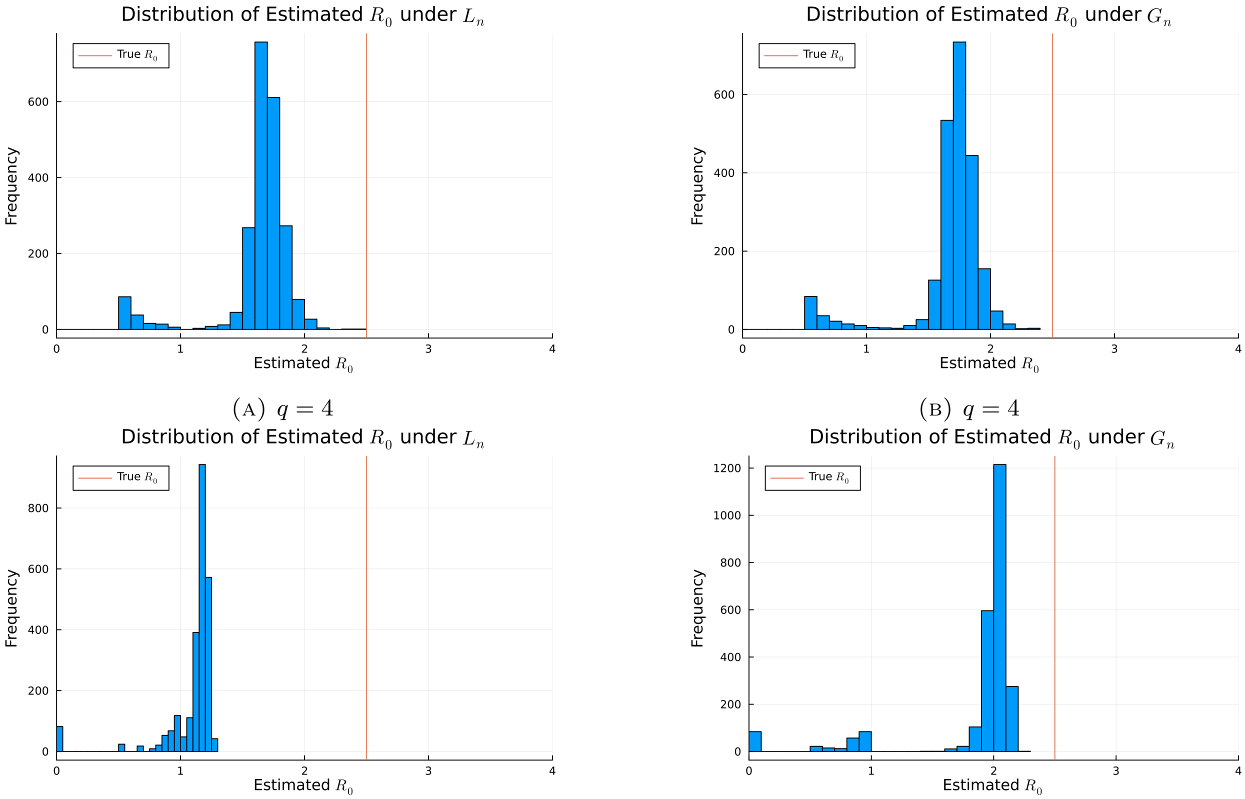


(A) $q = 4$

(B) $q = 2$

FIGURE B.4. Differences between $\hat{Y}_T(L_n)$ and $\hat{Y}_T(G_n)$ and the fitted values from the differential equation SIR model, for both $q = 4$ and $q = 2$.

As discussed above, Figure B.5 demonstrates that the fitted value of $\hat{\mathcal{R}}_0$ is typically below the true value of $\mathcal{R}_0 = 2.5$. In particular, with $q = 2$ and $L_n$, the estimation procedure dramatically underestimates the true value of $\mathcal{R}_0$. As discussed in the main text, this is because the estimation procedure does not use the micro-data of exactly which nodes are activated and when, as suggested in Proposition 1.



(A) $q = 4$

(B) $q = 4$

(C) $q = 2$

(D) $q = 2$

FIGURE B.5. Distribution of estimated $\hat{\mathcal{R}}_0$ across simulations when $L_n$ is based on $q = 4$. Note that the distribution of values sits below the true value of $\mathcal{R}_0 = 2.5$. Values very close to zero come from data where the epidemic stops after a very small number of activations.

In sum, a compartmental SIR model can, in many cases, be fit well looking backwards to a polynomial diffusion process. This fit is even better the higher the dimension of $L_n$, as it admits more expansive balls. But in all cases, the compartmental SIR estimates too rapid a diffusion that saturates and stabilizes too quickly: historical aggregate fits may be excellent and at the same time may serve as poor forecast tools.

B.6. **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 B.2. Results are shown in Figures B.6.

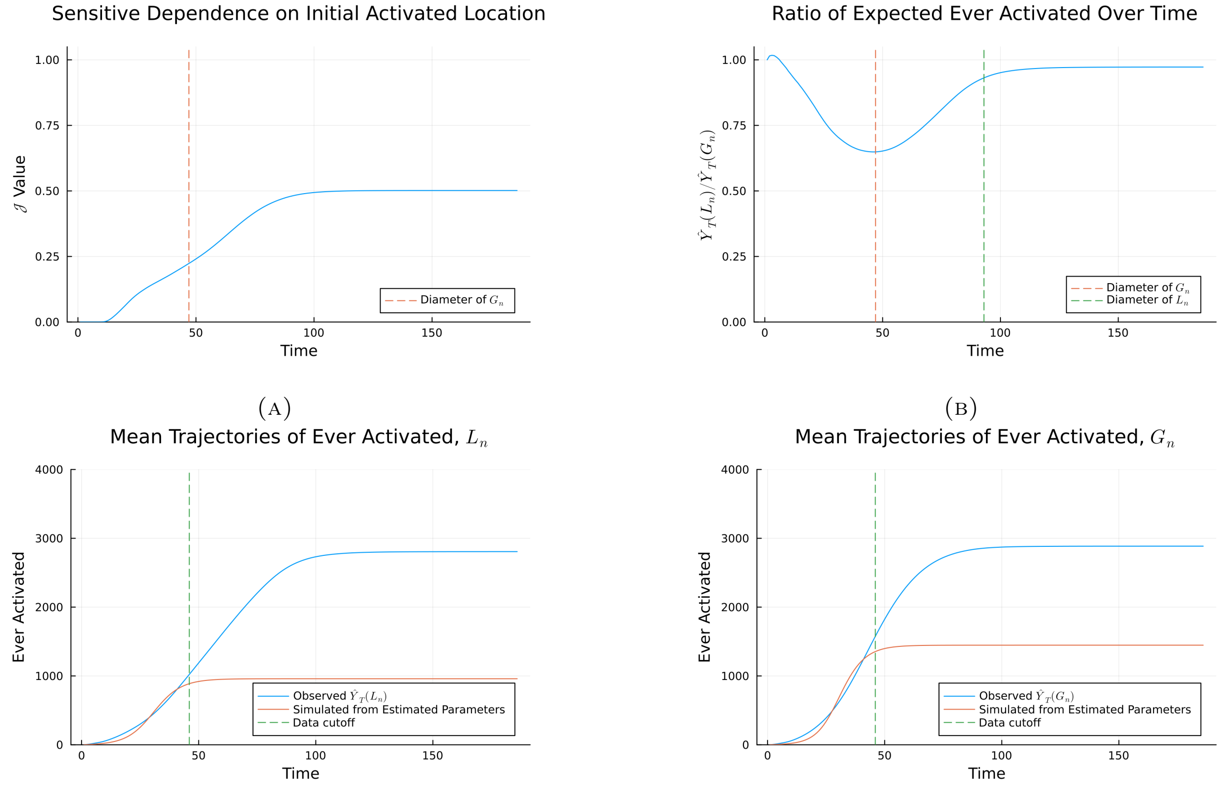
Table B.2. Graph statistics for $L_n$ generated with $q = 2$ and $G_n$ generated with $\beta_n = \frac{1}{100n}$

| 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.

As shown in Figure B.6, despite a much smaller value of $\beta_n$ forecasting issues persist. For sensitive dependence, $j_0$ is at distance 16 from $i_0$: this much larger distance comes from both the clustered nature of the graph, and the lack of i.i.d. links to connect disparate locations (due to the low value of $\beta_n$). Because there are so few links in $E_n$, due to the small value of $\beta_n$, the local neighborhood containing all $j_0$ is 7.13 percent of the graph, and only 10.90 percent of the neighborhood are candidate $j_0$. With this in mind, it is not surprising to see the process exhibit severe sensitive dependence on the seed location: at half of the diameter of $G_n$ ($T = 22$), the value of $\mathcal{J} = 0.09$ on average, indicating almost totally disjoint diffusion processes. For forecasting diffusion volume, the minimum value of $\hat{Y}_T(L_n)/\hat{Y}_T(G_n)$ is achieved at $T = 46$, taking a value of 0.649. This value is still lower than the case with $q = 4$ and $\beta_n = \frac{1}{10n}$ (which had a minimum of 0.780), showing the extreme sensitivity in the lower dimension. Note that over very short time ranges, the value of the ratio is slightly above 1 – this is a result of finite sample noise, with several diffusion processes on $L_n$ infecting a large number of nodes quickly, and a few processes on $G_n$ infecting very few nodes.

The third and fourth panels of Figure B.6 show the compartmental SIR fitting exercise. Here, the introduction of $E_n$ has less of an impact, as shown by the relative similarity between the results for $L_n$ and $G_n$. This result is not surprising, given the very small value of $\beta_n$. Recall that we fit the SIR model to $\hat{Y}_T(L_n)$ and $\hat{Y}_T(G_n)$, over the first 46 time steps (corresponding to $T/4$, equivalent to half of the diameter of $L_n$). In the fitting period, using $\hat{Y}_T(L_n)$, the average RMSE is 62.069, while in the next $T/4$ periods it is 1235.168 – a very similar set of values to the $q = 2$ case in the main text. With $\hat{Y}_T(G_n)$, the within sample average RMSE is 101.128, while in the next $T/4$ periods it is 1242.687. These values are much more similar to the $L_n$ case than the corresponding values for $q = 2$ in the main text – this is because there are many fewer additional links in $G_n$. Therefore, while the additional links increase the dimensionality of the diffusion process, the compartmental SIR model still gives a poor approximation. As further evidence, in both cases, the compartmental model dramatically underestimates the true value of $\mathcal{R}_0 = 2.5$: under $L_n$ it is estimated as 1.10, and under $G_n$ it is estimated as 1.21.



(A)

(B)

(C)

(D)

Figure B.6. Results with $q = 2$ and $\beta_n = \frac{1}{100n}$. Panel (A) shows the Jaccard index $\mathcal{J}$, while Panel (B) shows the ratio $\hat{Y}_T(L_n)/\hat{Y}_T(G_n)$. Panels (C) and (D) show the fitted values of a SIR differential equation model, fit to $\hat{Y}_T(L_n)$ and $\hat{Y}_T(G_n)$. Averages are taken over 2,500 Monte Carlo simulations.

## Appendix C. Empirical Example: Location Data from the COVID-19 Epidemic

We give a detailed description of the data processing procedures, along with additional results using a graph constructed from location data. We build a network using visitor flows based on cell phone location data, provided by SafeGraph (Kang et al., 2020). Our primary analysis studies the entirety of California and Nevada, with a small portion of Arizona included. Note that we only include areas in the United States. The region includes major cities including San Francisco, Los Angeles, and Las Vegas. We work with Census tracts as the unit of observation, which each contain approximately 4,000 people. Given privacy concerns, we focus on movement between tracts, rather than tracking individual people. We use tract-to-tract flows on March 1st, 2020. This date was before the WHO declared COVID-19 a pandemic, and before the United States government declared a national state of emergency. We construct graphs in the following manner. Fix a cutoff $c$. Then we take the following steps.

(1) For each pair of Census tracts $a$ and $b$, we construct the average flow between tracts by taking the average of the flow from $a$ to $b$ and the flow from $b$ to $a$. Call this value $f_{ab}$.
(2) Tracts $a$ and $b$ will be linked in the graph only if $f_{ab} > c$.

We choose $c$ based on the empirical distribution of $f_{ab}$, the flows between tracts. We refer to this procedure as "pruning." If the process results in a disconnected graph, we choose the largest connected subgraph. As before, we set $T$ as twice the diameter of $L_n$.

C.1. **Disease Process.** As with the simulated graphs, we fix $\mathcal{R}_0 = 2.5$. We then compute $p_n = \mathcal{R}_0/\bar{d}$, where $\bar{d}$ is the average degree in $L_n$. Note that in this case, the meaning of $\mathcal{R}_0$ is substantively different – because nodes now refer to Census tracts, infecting 2.5 nodes in the disease free state on average means infected 2.5 tracts on average.

C.2. **Errors Induced by Cutoff Choice.** We first study errors induced by choosing different cutoffs for pruning the graph. We construct $G_n$ by setting $c = 5$, which is at the 91st percentile of the empirical distribution of tract-to-tract flows. Then, we generate $L_n$ by choosing $c = 6$. Note that every link in $L_n$ will be in $G_n$, meaning that we can construct the implied error graph $E_n$.

We conduct the same three analyses that we did with the simulated graph. First, we study a version of Theorem 1, comparing the overlap between epidemics after perturbing the starting point. Second, we study a version of Theorem 2, comparing the expected number of infections on each graph. Finally, we consider the exercise of fitting a SIR differential equation model.

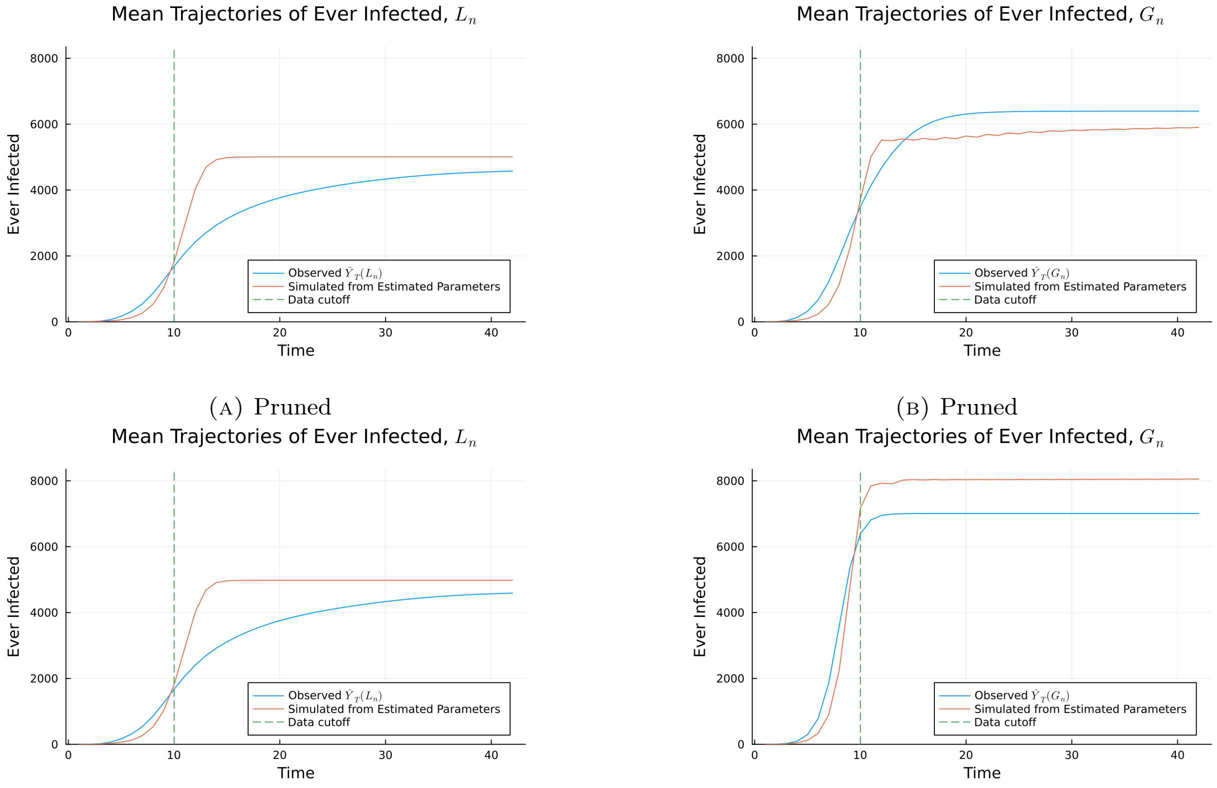
For the sake of brevity, we only note differences unique to this section when compared to the procedures discussed in Section B. When considering the simulation of Theorem 1, the only change is how $i_0$ is selected – we set $i_0$ to be the node with the highest degree in $G_n$. When considering the simulation of Theorem 2, the key change is that we hold $G_n$ fixed: it is generated once from the data. When we take expectations, they are taken only over the disease process only. Otherwise, the process is identical. The process of fitting a differential equation SIR model is exactly as before. In addition, we conduct simulations with $E_n$ taken to be an Erdos-Renyi random graph, rather than via the pruning procedure. In the main text, we set $\beta_n$ so that the i.i.d. errors generate the same expected volume of links as the pruning procedure. As an additional set of results, we set $\beta_n = \frac{1}{10n}$, to compare with the Monte Carlo simulations. Summary statistics of the resulting graphs are shown in Table C.1.

TABLE C.1. Graph statistics for $L_n$ and both hypothetical $G_n$s constructed from California, Nevada, and Arizona Census tract flow data

| Statistic | $L_n$ | $G_n^{92}$ | $G_n^{\beta}$ |
|---|---|---|---|
| Error Type | — | Pruned | IID |
| Diameter | 21.0 | 15.0 | 7.687 |
| Mean Degree | 12.962 | 15.486 | 16.172 |
| Min Degree | 1.0 | 1.0 | 1.839 |
| Max Degree | 298.0 | 329.0 | 301.148 |
| Mean Clustering Coefficient | 0.389 | 0.393 | 0.234 |
| Average Path Length | 7.253 | 5.866 | 4.03 |

Statistics for $G_n^{\beta}$ with i.i.d. errors are averaged over 2,500 draws.

C.3. **Additional Results.** We again estimate the compartmental SIR model using the simulated epidemics above. This process is identical to the procedure conducted in Section 6. One pattern of note is that the model fit to $\hat{Y}_T(G_n)$ generated from the pruning procedure underestimates the average number of infections, while the model fit to $\hat{Y}_T(L_n)$ overestimates.



(A) Pruned

(B) Pruned

(C) I.I.D.

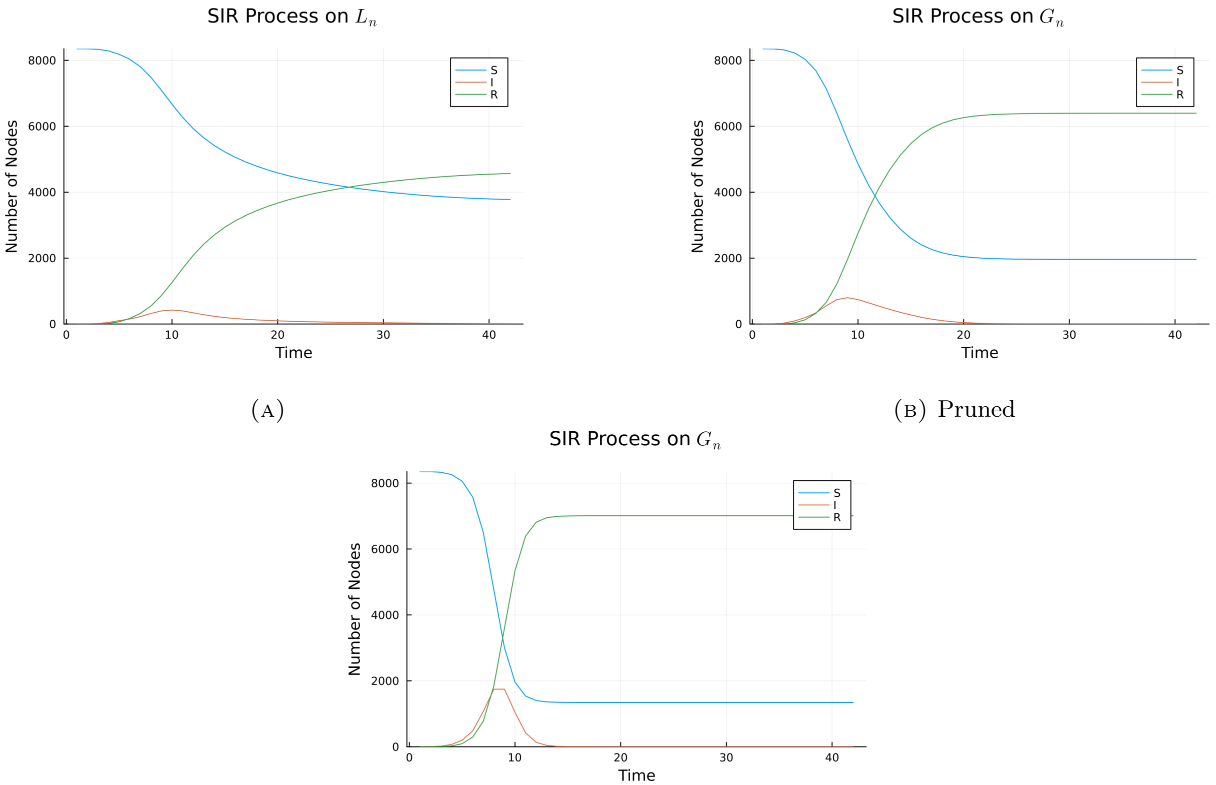
(D) I.I.D.

FIGURE C.1. A comparison of the mean ever infected under the true network SIR model and the estimated trajectory from the differential equations model. Here, $L_n$ is generated from location flow data in California, Nevada, and a portion of Arizona. Panel (A) and (B) use the pruning procedure, while (C) and (D) have i.i.d. links. Panel (A) shows simulations when $\hat{Y}_T(L_n)$ is used as the data generating process, while Panel (B) shows when $\hat{Y}_T(G_n)$ is used. The data cutoff is at $T/4$. Before this point, the SIR model is fit to the data via the generalized method of moments. After this point, we run the differential equation model forward to see how it performs out of sample.

In the estimation period before $T/4$, the RMSE for $\hat{Y}_T(L_n)$ is 202.98, while in the next $T/4$ periods it is 1953.41. When fit to $\hat{Y}_T(G_n^{93})$, the RMSE in the first $T/4$ periods is 452.09, while in the next $T/4$ periods it is 1320.60. Notably, the model has a much better fit out of sample for $G_n^{93}$. For the i.i.d. errors on $G_n^{\beta}$, the results are similar. In the estimation period, the RMSE fitted to $\hat{Y}_T(L_n)$ is 200.541, while in the next $T/4$ periods it is 1944.63. When fit to $\hat{Y}_T(G_n)$, the RMSE in the first $T/4$ periods is 700.93, while in the next $T/4$ periods it is 1095.14.

We then show a set of additional figures, corresponding to the simulations from the main text. We first disaggregate the simulated diffusion processes into a standard SIR framework, as shown in Figure C.2. Second, we show the distribution of estimated $\hat{\mathcal{R}}_0$ across simulations in Figure C.3. Figure C.2 demonstrates that with i.i.d. errors, the infection profile is relatively sharp, as the epidemic quickly expands to cover the whole graph during the intermediate range of $T$.

C.4. **Lower Rates of I.I.D. Errors.** 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



(A)

(B) Pruned

(C) I.I.D.

FIGURE C.2. Trajectories of $\hat{Y}_T(L_n)$ and $\hat{Y}_T(G_n)$ disaggregated into the standard SIR curves for $L_n$ and $G_n$ for each scenario. Note that the $L_n$ specifications are identical, as it is exactly the same graph.

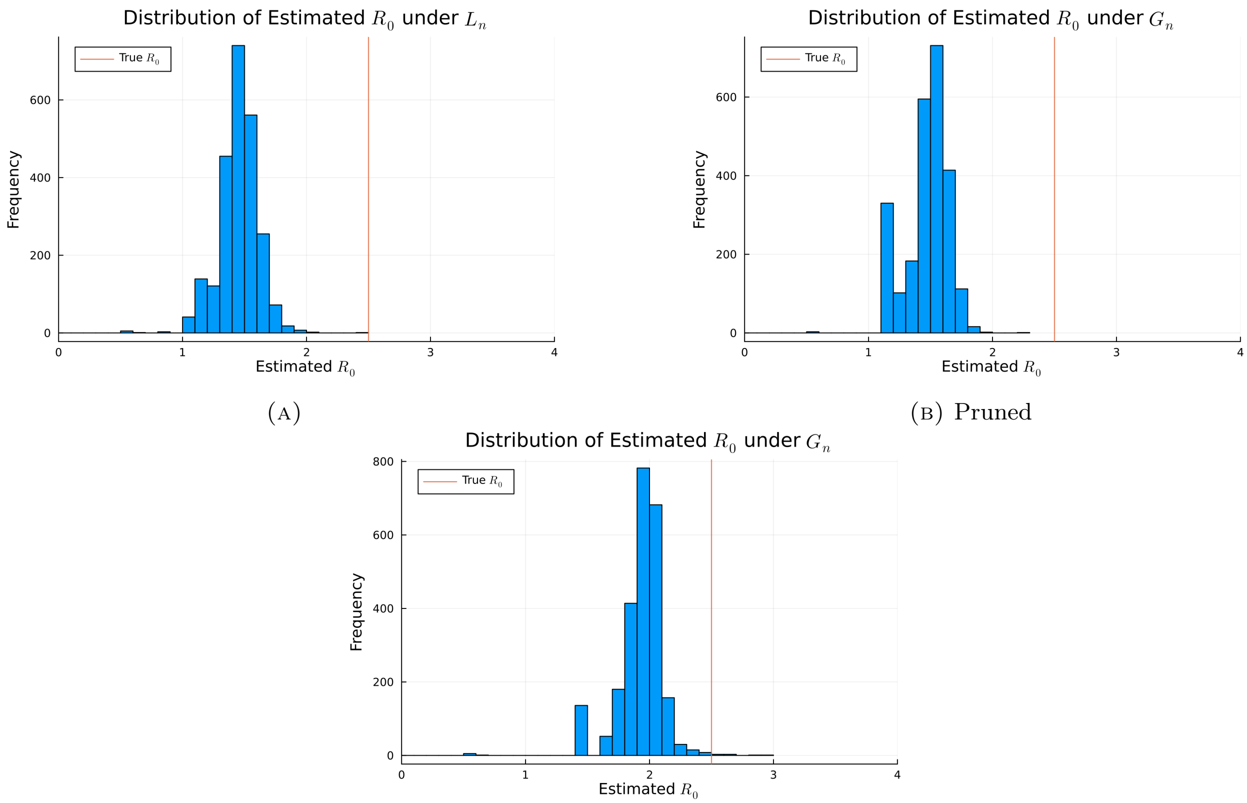
Table C.2, again for $L_n$ and the average statistics for $G_n$ over 2,500 draws of $E_n$. Compared to $G_n$ in the main text (in Table C.1), note that the change in degree, clustering, and average path length are all much smaller, as $E_n$ is much more sparse in this case.

TABLE C.2. 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 | 16.874 |
| Mean Degree | 12.962 | 13.062 |
| Min Degree | 1.0 | 1.0 |
| Max Degree | 298.0 | 298.106 |
| Mean Clustering Coefficient | 0.388 | 0.38 |
| Average Path Length | 7.295 | 6.116 |

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

Results are shown in Figure C.4. 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. Of the neighborhood, 89.55 percent of the nodes are candidates for $j_0$. Halfway to the diameter of $G_n$, the value



(A)

(B) Pruned

(C) I.I.D.

FIGURE C.3. The distribution of values of $\hat{\mathcal{R}}_0$ estimated when fitting the compartmental SIR model to the COIVD-19 travel data.

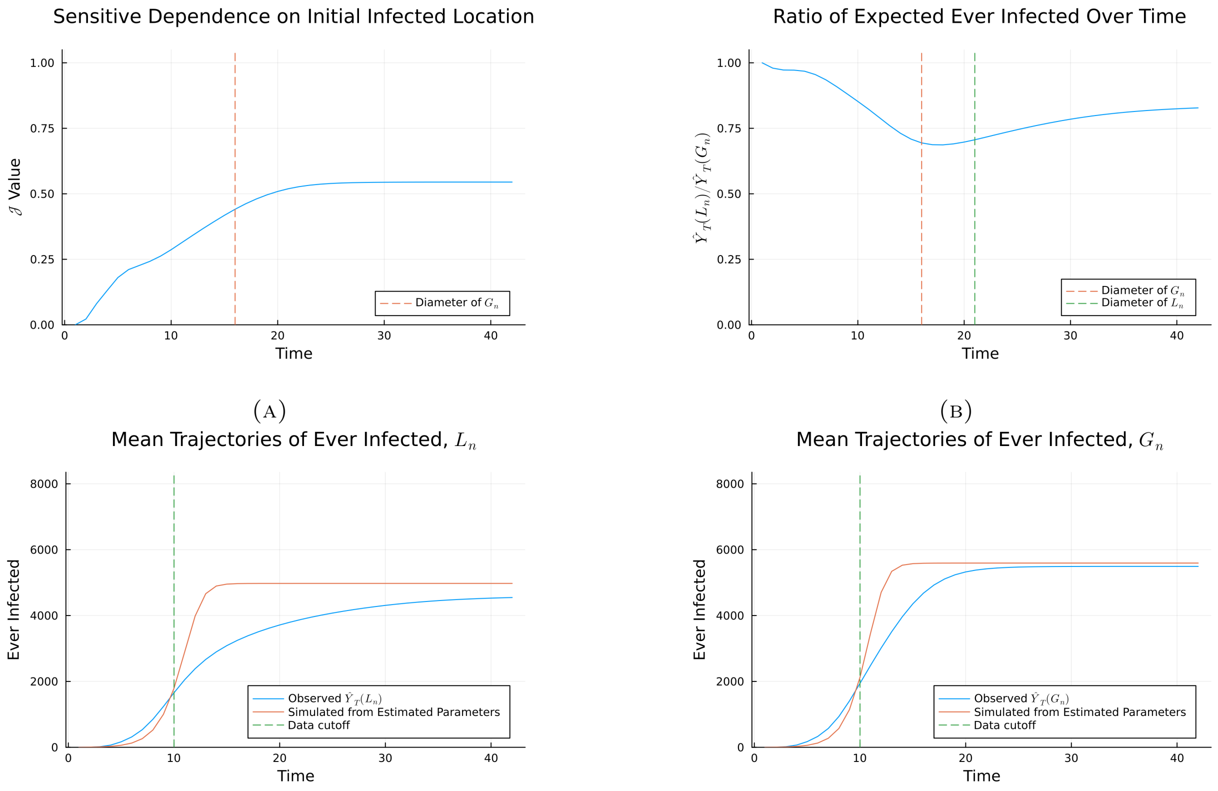
of the average Jaccard index is 0.24, indicating largely distinct epidemics. The top right panel shows the simulation of Theorem 2. Note that in this case, the minimum ratio of $\hat{Y}_T(L_n)/\hat{Y}_T(G_n)$ is achieved at $T = 18$ and takes the value 0.686. This value is much larger than the values from the main text with either the pruned or i.i.d. errors, and comparable to the values with the same level of $\beta_n$ and graph dimension $q = 4$ in the Monte Carlo simulations.

The third and fourth panels of Figure C.4 show the fitted compartmental SIR models, relative to $\hat{Y}_T(L_n)$ and $\hat{Y}_T(G_n)$. As before, the compartmental model underestimates the true $\mathcal{R}_0 = 2.5$: under $\hat{Y}_T(L_n)$, it estimates a value of 1.40, and under $\hat{Y}_T(G_n)$ estimates a value of 1.49. In the first $T/4$ periods, in sample, the average RMSE under $\hat{Y}_T(L_n)$ is 198.96. In the next $T/4$ periods, it is 1,966.58. Under $\hat{Y}_T(G_n)$, in sample, the average RMSE is 222.11, whereas in the next $T/4$ periods it is 1389.65. Similar to the Monte Carlo exercise, we see that the additional links in $E_n$ help increase the dimensionality of the epidemic, leading to a better fit with the exponential compartmental model.

## APPENDIX D. EMPIRICAL EXAMPLE: DIFFUSION IN MOBILE PHONE MARKETING

We use data from Banerjee et al. (2019) as an additional empirical example of our diffusion results. We build 69 separate village graphs, by composing networks based on survey data from Karnataka, India. We have a number of directed networks:

(1) Relative
(2) Give advice: does the household $i$ give advice to household $j$



(A)

(B)

(C)

(D)

FIGURE C.4. 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 $\mathcal{J}$. Panels (C) and (D) show the fitted values of a SIR differential equation model, fit to $\hat{Y}_T(L_n)$ and $\hat{Y}_T(G_n)$. Averages are taken over 2,500 Monte Carlo simulations.

(3) Seek advice: does household $i$ get advice from household $j$
(4) Go to visit: does household $i$ visit household $j$ in free time
(5) Come to visit: does household $i$ come visit household $j$ in free time
(6) Borrow: does $i$ borrow kerosene or rice from household $j$
(7) Lend: does $i$ lend kerosene or rice to household $j$

To construct a set of undirected networks for each village, we take the union of these seven networks. Links are assumed to be undirected, and the network is made symmetric. This network data comes from a sequence of studies conducted in Karnataka, India. We use the 2012 data in our setting, the more recent of two waves of data collection. Graph statistics are shown in Table D.1.

## APPENDIX E. EMPIRICAL EXAMPLE: PEER EFFECTS IN INSURANCE

We use data from Cai et al. (2015) to investigate an example with peer effects in a diffusion setting. In order to encourage weather insurance, a valuable product with low takeup in rural China, the researchers conducted two waves of information sessions.

To construct network data, we use the list of directed links given in their data along with additional survey data. We drop some households who are listed in the network data but not in the additional survey data – we assume that this is a result of attrition between the surveys. We then transform the directed network in

Table D.1. Average village graph information from Banerjee et al. (2019).

| Statistic | $L_n$ | $G_n$ |
|---|---|---|
| Nodes | 196.072 | 196.072 |
| Diameter | 7.087 | 6.787 |
| Mean Degree | 6.541 | 6.849 |
| Min Degree | 1.0 | 1.004 |
| Max Degree | 25.71 | 26.219 |
| Mean Clustering Coefficient | 0.228 | 0.199 |
| Average Path Length | 3.303 | 3.168 |

For $L_n$, averages are taken across the 69 villages in our sample. For $G_n$, averages are taken across the 69 villages and 2,500 draws of $E_n$, where $E_n$ is generated with $\beta_n = \frac{1}{2n_v}$ in each village separately, where $n_v$ is the number of households in the village.

each village into an undirected network: if household $i$ lists household $j$ as a friend, or vice versa, we link $i$ to $j$.

We use the same definition of treatment as in Cai et al. (2015). A household is considered to be treated if they participate in an intensive information session in the first wave of the experiment. We then compute diffusion exposure using these households as seeds. When we estimate the effect of diffusion exposure, we include only households that did not participate in the first wave of the information sessions. This procedure is consistent with the prior research.

In addition, we include a number of controls to be in line with the original paper. We control for the head of household gender, age, education, and area of rice production. In addition, following the approach in Cai et al. (2015), we control for degree to address potential concerns about selection on household sociability. Finally, we include village fixed effects. Tables E.1 and E.2 report graph summary statistics for all values of $k$ for the Monte Carlo simulations conducted in the main text.

Table E.1. Average graph statistics from Cai et al. (2015)

| Graph Statistic | Value |
|---|---:|
| Nodes | 104.30 |
| Min Degree | 0.40 |
| Max Degree | 15.79 |
| Mean Degree | 6.51 |
| Components | 5.60 |
| Average Path Length | 3.59 |
| Diameter | 8.06 |
| Local Clustering | 0.30 |
| Exposure | 0.99 |

Averages are taken over the 47 villages in the data. When there are multiple components, paths of infinite length (when nodes are disconnected from one another) are ignored. Mean exposure is computed before standardizing to have mean zero and standard deviation one, as we do in the regressions.

## Appendix F. Additional Theoretical Results

F.1. **Decaying Diffusion in the Polynomial Case.** In what follows, we set $\delta_n = 1$ for ease of exposition. Similar results go through with vanishing $\delta_n$ at a sufficiently slow rate and adjustments to the bounds on timing.

Table E.2. Graph statistics for the average graph $L_n$ generated by dropping links with i.i.d. probability $\beta_n = \frac{1}{kn_v}$ in each village

| k | MinDeg | MaxDeg | MeanDeg | Comp. | PathLen. | Diam. | Clus. | Exposure |
|---:|---:|---:|---:|---:|---:|---:|---:|---:|
| - | 0.38 | 13.32 | 5.60 | 5.60 | 3.59 | 8.06 | 0.30 | 1.10 |
| 15.00 | 0.37 | 13.19 | 5.54 | 5.63 | 3.61 | 8.10 | 0.29 | 1.10 |
| 14.00 | 0.37 | 13.19 | 5.53 | 5.64 | 3.61 | 8.10 | 0.29 | 1.10 |
| 13.00 | 0.37 | 13.17 | 5.53 | 5.64 | 3.61 | 8.10 | 0.29 | 1.10 |
| 12.00 | 0.37 | 13.16 | 5.52 | 5.64 | 3.61 | 8.11 | 0.29 | 1.10 |
| 11.00 | 0.37 | 13.15 | 5.51 | 5.65 | 3.61 | 8.11 | 0.29 | 1.10 |
| 10.00 | 0.37 | 13.13 | 5.51 | 5.65 | 3.62 | 8.12 | 0.29 | 1.10 |
| 9.00 | 0.37 | 13.11 | 5.49 | 5.66 | 3.62 | 8.12 | 0.29 | 1.10 |
| 8.00 | 0.36 | 13.09 | 5.48 | 5.67 | 3.62 | 8.13 | 0.29 | 1.10 |
| 7.00 | 0.36 | 13.05 | 5.46 | 5.68 | 3.62 | 8.14 | 0.29 | 1.10 |
| 6.00 | 0.36 | 13.01 | 5.44 | 5.69 | 3.63 | 8.15 | 0.29 | 1.09 |
| 5.00 | 0.35 | 12.95 | 5.40 | 5.71 | 3.64 | 8.17 | 0.28 | 1.09 |

“Comp." stands for the number of components. “PathLen.” stands for path length. “Diam.” stands for diameter. “Clus.” stands for the clustering coefficient.

**Assumption F.1** (Polynomial Diffusion Process)**.** *For some constant $q > 1$ and all discrete-time $t$, $\mathcal{E}_t = \Theta(t^{q+1})$ and $\mathcal{S}_t = \mathcal{E}_t - \mathcal{E}_{t-1} = \Theta(t^q)$. Furthermore, let $p_n(T) = \frac{p_{n,0}}{T^\lambda}$ for some constant $0 < \lambda < q$ and $p_{n,0} \in \left(\left(\frac{1}{\log n}\right)^{\frac{q}{2q+2}}, 1\right]$.*

We assume that diffusion decays at a polynomial rate over time, governed by the constant $\lambda$. It follows that we only have expected diffusion through links in $E_n$ if $\lambda < q$. In the case where $p_n(T)$ were to decay exponentially quickly, it would then follow we still would not have expected diffusion through $E_n$. With fast decay on $p_n(T)$, the missing links do not have a large impact because the diffusion process just dies before hitting regions that cause lots of damage.

We note that the graph classes that are allowed under the homogenous $p_n$ are still valid here, with sufficiently slow decay of $p_n$. For example, again consider a latent space network where nodes form links locally in a Euclidean space with dimension $q$. Since volumes in Euclidean space expand at a polynomial rate and for $\lambda < q$, this ensures that Assumption F.1 will be satisfied.

**Assumption F.2** (Forecast Period)**.** *We impose that the sequence $T_n$ has for each $n$, $T_n \in [\underline{T}_n, \overline{T}_n]$ where the following holds:*

(1) $\overline{T}_n = \min\left\{n^{\frac{1}{q+1}}, \left(\frac{n}{p_{n,0}}\right)^{\frac{1}{q-\lambda}}\right\}$

(2) $\underline{T}_n = (p_{n,0} \log n)^{\frac{1}{q+\lambda+2}}$.

Again, the assumption is very close to that of the main text, adapting the constants to deal with the decaying diffusion rates. We can note that the time frame considered will generally start earlier, but also potentially end earlier.

**Assumption F.3.** $\beta_n \in \left(\frac{1}{np_{n,0}T^{q-\lambda}}, \frac{1}{n}\right)$.

Compared to the homogeneous diffusion rate case where we assume $\beta_n \in \left(\frac{1}{np_{n,0}T^q}, \frac{1}{n}\right)$, with decaying $p_n(T)$ we impose a larger missing link rate. Under these conditions, a similar result to Theorem 2 holds.

**Theorem F.1.** *Under Assumptions F.1, F.2, and F.3, as $n \to \infty$, $\frac{\hat{Y}_T(L_n)}{\hat{Y}_T(G_n)} \to 0$.*

This result forms a direct analogue of Theorem 2 with a decaying diffusion rate. The proof is omitted as it proceeds in an identical manner.