Non-Robustness of Diffusion Estimates on Networks with Measurement Error

Tyler H. McCormick
University of Washington

IMSI Networks

January 14, 2026

Collaborators

- Arun G. Chandrasekhar, Samuel Thau (Stanford econ)
- Paul Goldsmith-Pinkham (Yale SoM)
- Jerry Wei (formerly UW)

This talk in one slide

Measurement and measurement error are important, and if you deal with it in a hacky way, then your statistical inference can be really bad.

–David Dunson, actual Bayesian

This talk in one slide

Measurement and measurement error are important, and if you deal with it in a hacky way, then your statistical inference can be really bad.

–David Dunson, actual Bayesian

Mismeasurment and measurment error are important and if you deal with it in **really antagonistic** *way, then your* **forecasts** *can be really bad.*

–Tyler, Bayesian non-realist

In contrast to Karl....

So a "Bayesian non-realist" would be someone who:

- Uses the full Bayesian machinery—priors, posteriors, model comparison
- But treats the model as a useful fiction rather than a claim about truth
- Cares about predictive performance and decision quality, not whether the latent factors "really exist"
- Is comfortable saying "this model works well" without saying "this model is true"

Honestly, it fits pretty well with some of the themes in your work—the fundamental limits piece on observing proxies rather than true phenomena has a distinctly non-realist flavor (we can't access the real thing, only its shadows). Your factor analysis work on the physics data also has this quality: you're extracting useful structure and labeling it with physical concepts, but the *representation* is a tool for interpretation, not a claim that the autoencoder has discovered the universe's true ontology.

Did he mean it as a compliment, a critique, or just an observation?

Models of Diffusion on Networks

Researchers/ policymakers studying the spread of ideas, technology, or disease often estimate models of diffusion using network data on how individuals interact.

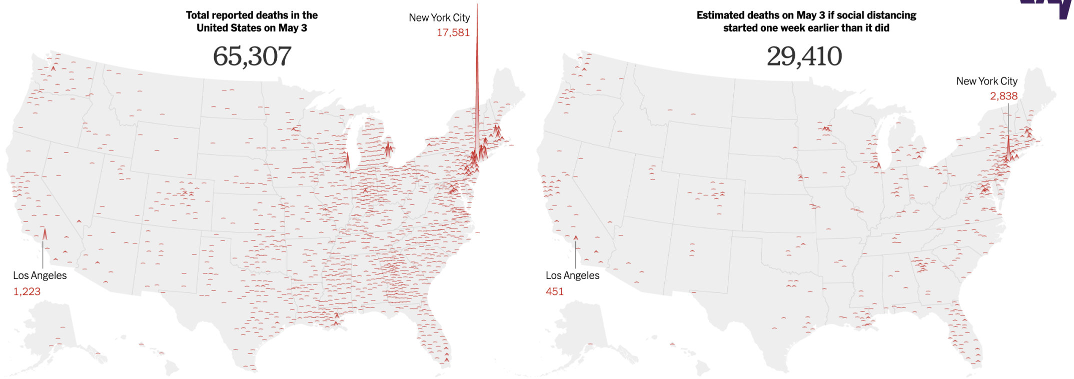
Examples:

1. quantifying the extent of illness or technology take-up;
2. summarizing diffusion dynamics (e.g., $\mathcal{R}_0$ of a disease);
3. targeting interventions

   e.g., where to seed new information to maximize spread, where to lockdown to prevent spread;
4. estimating counterfactuals

   e.g., in estimates of peer effects, as we show in an empirical example.

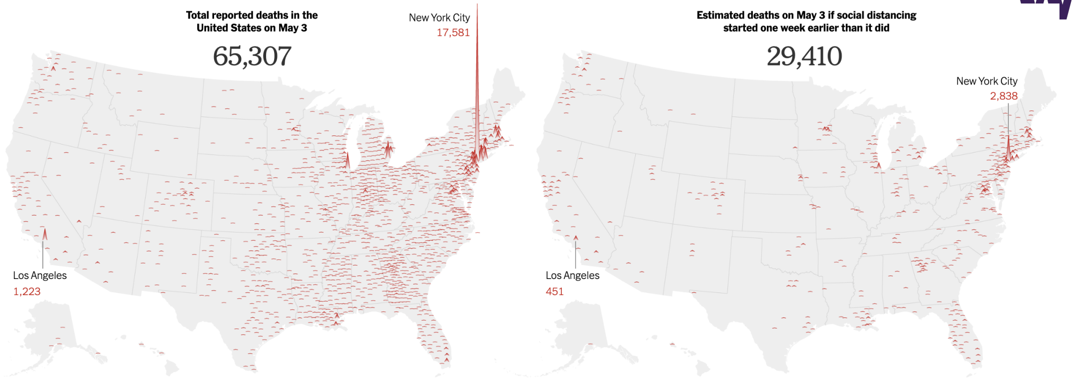
Counterfactual predictions of Covid-19 infections + deaths



By Lazaro Gamio • Source: "Differential Effects of Intervention Timing on COVID-19 Spread in the United States," by Sen Pei, Sasikiran Kandula and Jeffrey Shaman, Columbia University

Modelers find that tens of thousands of U.S. deaths could have been prevented.

Counterfactual predictions of Covid-19 infections + deaths



By Lazaro Gamio • Source: "Differential Effects of Intervention Timing on COVID-19 Spread in the United States," by Sen Pei, Sasikiran Kandula and Jeffrey Shaman, Columbia University

Modelers find that tens of thousands of U.S. deaths could have been prevented.

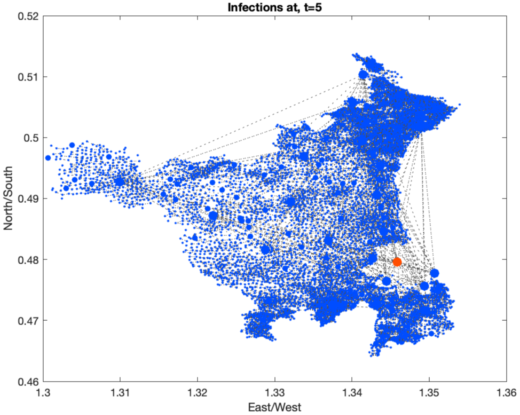
How did they model this?

six | seven

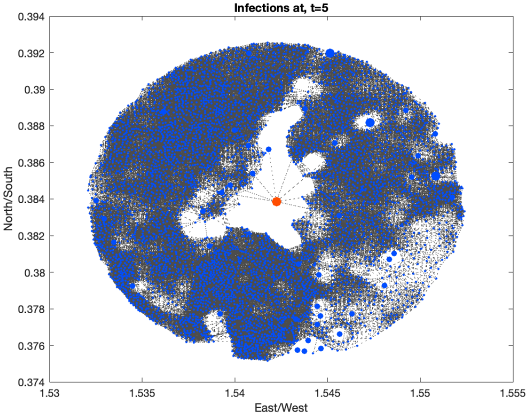
AT THE EDGEWATER

Rich Covid-19 SIERD Model

Haryana

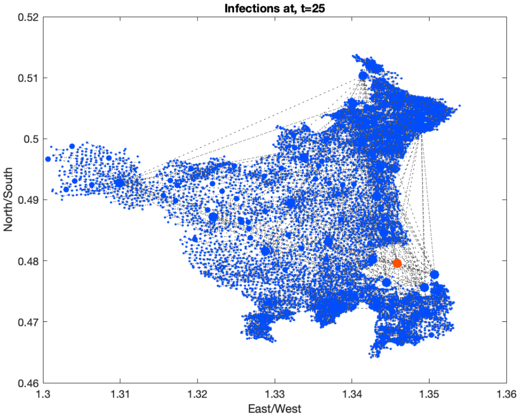


West Bengal (zoomed on Kolkata)

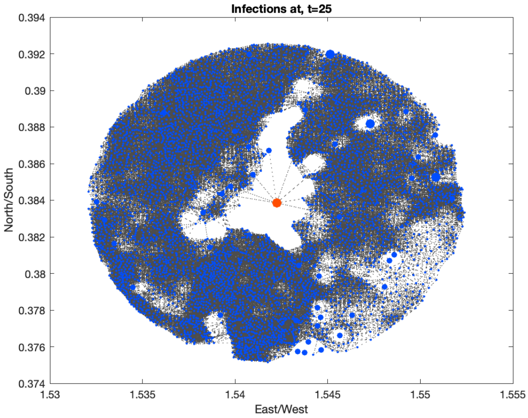


Rich Covid-19 SIERD Model

Haryana

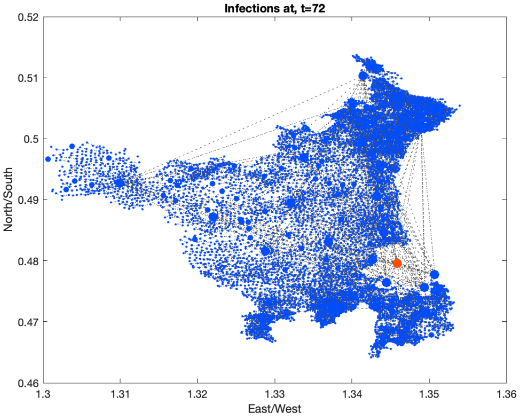


West Bengal (zoomed on Kolkata)

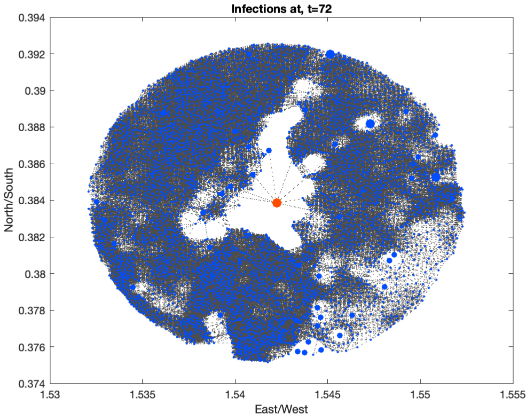


Rich Covid-19 SIERD Model

Haryana

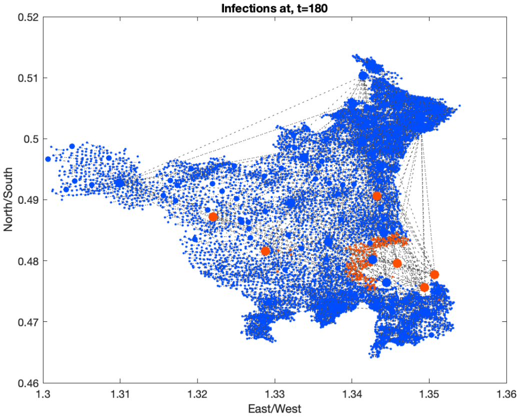


West Bengal (zoomed on Kolkata)

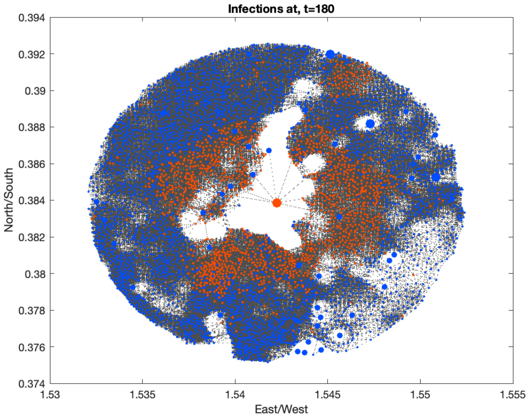


Rich Covid-19 SIERD Model

Haryana

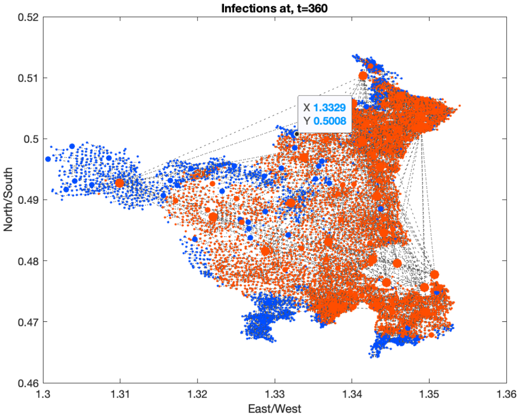


West Bengal (zoomed on Kolkata)

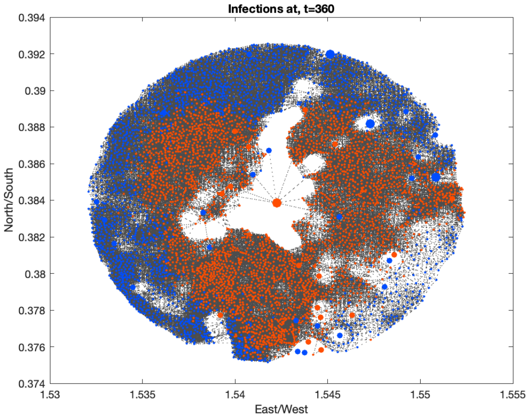


Rich Covid-19 SIERD Model

Haryana



West Bengal (zoomed on Kolkata)



Network Mismeasurements

Several reasons:

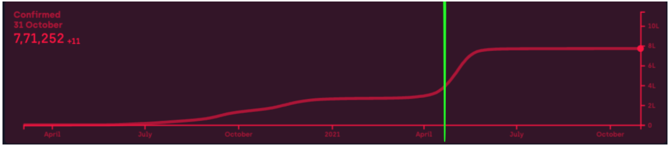
- Who is seeded? Identity of i_0...
- The sampling process for the network is imperfect.
  - surveys, geo, mobility, online,..
  - philosophical: e.g., referrals – cricket links? advice? kin? RoSCAs?
- It may be that a rich snapshot of a network does not capture the relevant links for diffusion by the time the process reaches an individual.

$$G^t_{i,\cdot} \neq G^0_{i,\cdot}$$

- Many analyses using empirical data do some amount of aggregation into groups with measured amounts of interaction.

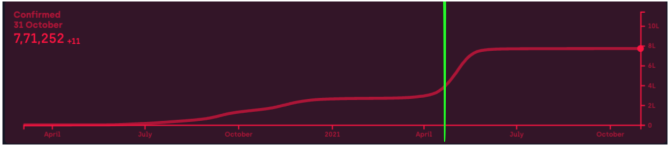
Time Horizon: Medium Run

- (a) day 1 uninteresting; (b) in the long run diffuses to giant component
- scope for intervention is in the "sleeve" in the prior to the explosion



Time Horizon: Medium Run

- (a) day 1 uninteresting; (b) in the long run diffuses to giant component
- scope for intervention is in the “sleeve” in the prior to the explosion



network structure (shape of diffusion) intimately related to time

- Joe Schmoe: slow spread
  - time to respond
- Taylor Swift: info diffuses to new heights
  - too late to respond

Main Results

1. known network G, very locally perturbed seed i_0:
   - predictions of **where** diffusion goes is very sensitive to *local* uncertainty of i_0

Main Results

1. known network G, very locally perturbed seed i_0:
   - predictions of **where** diffusion goes is very sensitive to *local* uncertainty of i_0
2. minuscule imperfections in knowledge of G, know i_0:
   - **counts** grossly under-estimated w/ even vanishingly small missing links;

Main Results

1. known network G, very locally perturbed seed i_0:
   - predictions of **where** diffusion goes is very sensitive to *local* uncertainty of i_0
2. minuscule imperfections in knowledge of G, know i_0:
   - **counts** grossly under-estimated w/ even vanishingly small missing links;
3. in this regimes, aggregated quantities often ok:
   - e.g., transmission prob., $\mathcal{R}_0$

Main Results

1. known network G, very locally perturbed seed i_0:
   - predictions of **where** diffusion goes is very sensitive to *local* uncertainty of i_0
2. minuscule imperfections in knowledge of G, know i_0:
   - **counts** grossly under-estimated w/ even vanishingly small missing links;
3. in this regimes, aggregated quantities often ok:
   - e.g., transmission prob., $\mathcal{R}_0$
4. many (practical) data augmentation approaches ineffectual:
   - trivial size of measurement error makes it hard to estimate w/ extra data collection
   - and realistic testing protocols will be behind the curve $\implies$ reseeding elsewhere

1. Environment

General Setup

Society = seq. of undirected, unweighted graph G_n

$G_n := L_n \cup E_n$, where L_n is base with min. degree d_L, and the links in E_n are Ber$(\beta_{ij,n})$.

- L_n perfectly observed by statistician
- E_n unobserved
  - e.g., sampling, compartmental smoothing, network shifting over time, ...
- $d_E / d_L \to 0$ (we assume much stronger vanishingness of E_n)
  - E_n *exceedingly* sparse..

Diffusion process: standard SIR on G_n with i.i.d. passing probability p_n.

The Diffusion Process

Assumption 1.

1. For some constant $q > 1$ and $t \in \mathbb{N}$, $|t\text{-radius ball in } L_n| = \Theta\left(t^{q+1}\right)$
2. Diffusion from the seed i_0 is super-critical on L_n, and each subset of a ball around i_0 is hit with positive probability

1. Polynomial growth (lattices, RGGs in Euclidean space, etc)
2. Diffusion is balanced across the graph $\Rightarrow$ no bottlenecks, $p_n > 0$

Implies the ever-activated set: $\mathcal{E}_t := \mathbb{E}\left|\{j \mid j \text{ ever activated by the diffusion on } L_n\}\right|$.
Is $\Theta(t^{q+1})$ as well. Expected increment of new activations is $\Theta(t^q)$

Policy Relevant Time Period

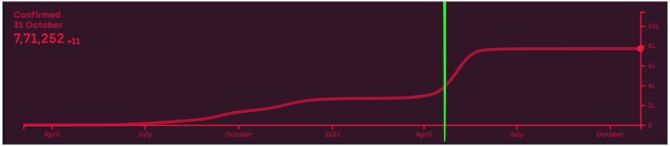
Assumption 2. (Time Period of Focus)

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) $\underline{T}_n = \omega(1)$

1. $\overline{T}_n$: the diffusion has not reached the edge of the graph
   - more expansive ($q \uparrow$), the earlier medium run ends
2. $\underline{T}_n$: the party has to get started...



Missing Mechanism

Assumption 3. (Missingness Pattern) Links in $E_n \sim \text{Bern}(\beta_n)$, with $\beta_n = \omega\left(\frac{1}{p_n n \underline{T}_n^q}\right), O\left(\frac{1}{n}\right)$

Comments:

1. Live in the interesting case where E_n has no giant component

Missing Mechanism

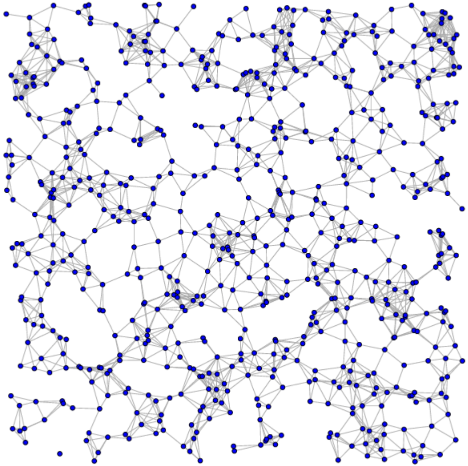
Assumption 3. (Missingness Pattern) Links in $E_n \sim \text{Bern}(\beta_n)$, with

$$\beta_n = \omega\left(\frac{1}{p_n n \underline{T}_n^q}\right), O\left(\frac{1}{n}\right)$$

Comments:

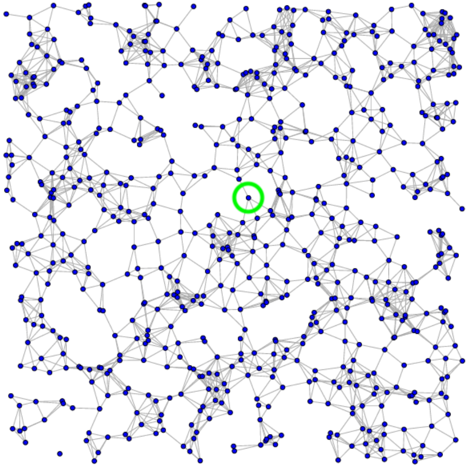
1. Live in the interesting case where E_n has no giant component
2. Can generalize to non-iid links: each node can only link to a vanishing fraction of the graph, heterogeneous by node
3. Key condition: errors are not too localized within a t-ball of the seed

Missing Mechanism



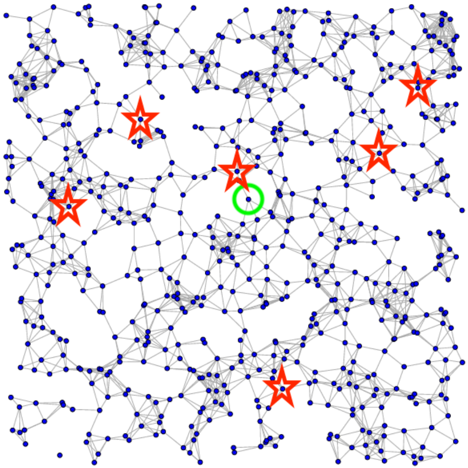
stylized L_n

Missing Mechanism



seed i_0

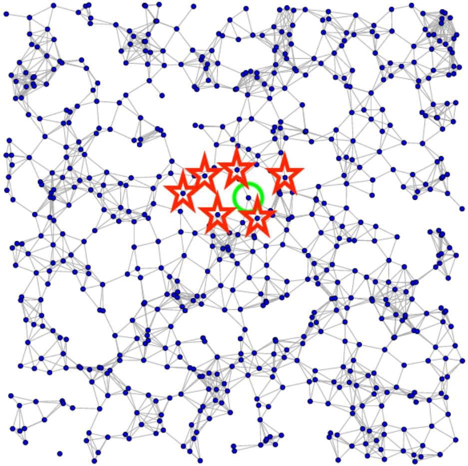
Missing Mechanism



global support:

- i_0 can only make links in E_n with ★
- but reside anywhere

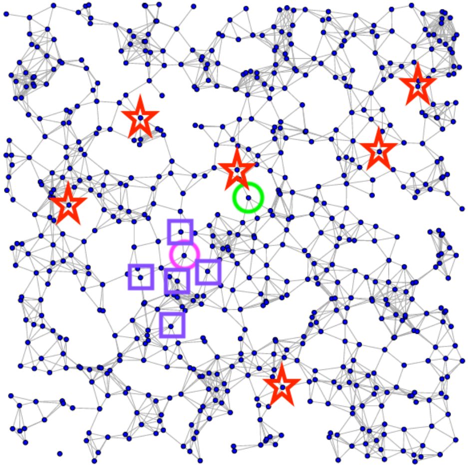
Missing Mechanism



local support:

- $\mathbf{i_0}$ can only make links in E_n with ★
- local in L_n

Missing Mechanism



two seeds $\mathbf{i_0}$ and j_0:

- $\mathbf{i_0}$ can only make links with ★
- j_0 can only make links with □

Sense of Scale

California: pop 38.9 million

- $q = 2$: upper bound 11 months
- $q = 3$: 3 months
- structures w/ rare links, local in L_n,

Haryana: pop 25.4 million

- $q = 2$: upper bound 10 months
- $q = 3$: 2.4 months
- structures w/ rare links, local in L_n,

2. Sensitive Dependence to Seed Set

Some Notation

- $I_P(k, T)$: the set of ever activated nodes from a diffusion process starting at k after T time steps
- Consider some starting point i_0
- We find the set of js that are (a) local to i_0; (b) local to some other k; (c) k can't be reached by i_0 in T periods
- For an alternative seed j_0, track overlap with a Jaccard index:

$$\Delta_n(i_0, j_0) := \frac{|I_P(i_0, T) \cap I_P(j_0, T)|}{|I_P(i_0, T) \cup I_P(j_0, T)|}.$$

Sensitive Dependence

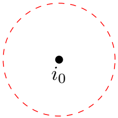
Theorem 1. Let Assumptions 1-3 hold and i_0 be a seed.

We find a set of alternative seeds J_{i_0} such that with positive probability (over (P_n, E_n)):

1. a non-vanishing share belongs to J_{i_0}: $|J_{i_0}|/|B^L_{i_0}| > c$
2. if we counterfactually seed $j_0 \in J_{i_0}$, a agreement is bounded from above

$$\Delta(i_0, j_0) < c' < 1.$$

3. many disjoint catchment areas form



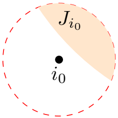
Sensitive Dependence

Theorem 1. Let Assumptions 1-3 hold and i_0 be a seed. We find a set of alternative seeds J_{i_0} such that with positive probability (over (P_n, E_n)):

1. a non-vanishing share belongs to J_{i_0}: $|J_{i_0}|/|B^L_{i_0}| > c$
2. if we counterfactually seed $j_0 \in J_{i_0}$, a agreement is bounded from above

$$\Delta(i_0, j_0) < c' < 1.$$

3. many disjoint catchment areas form



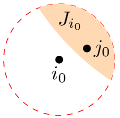
Sensitive Dependence

Theorem 1. Let Assumptions 1-3 hold and i_0 be a seed. We find a set of alternative seeds J_{i_0} such that with positive probability (over (P_n, E_n)):

1. a non-vanishing share belongs to J_{i_0}: $|J_{i_0}|/|B^L_{i_0}| > c$
2. if we counterfactually seed $j_0 \in J_{i_0}$, a agreement is bounded from above

$$\Delta(i_0, j_0) < c' < 1.$$

3. many disjoint catchment areas form



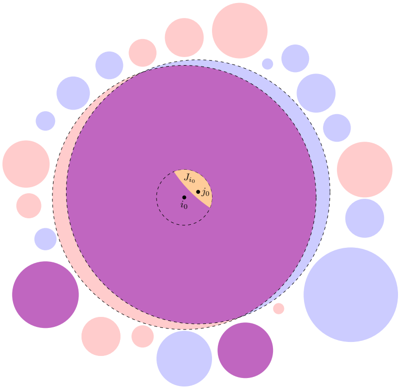
Sensitive Dependence

Theorem 1. Let Assumptions 1-3 hold and i_0 be a seed. We find a set of alternative seeds J_{i_0} such that with positive probability (over (P_n, E_n)):

1. a non-vanishing share belongs to J_{i_0}: $|J_{i_0}|/|B^L_{i_0}| > c$
2. if we counterfactually seed $j_0 \in J_{i_0}$, a agreement is bounded from above

$$\Delta(i_0, j_0) < c' < 1.$$

3. many disjoint catchment areas form



Discs

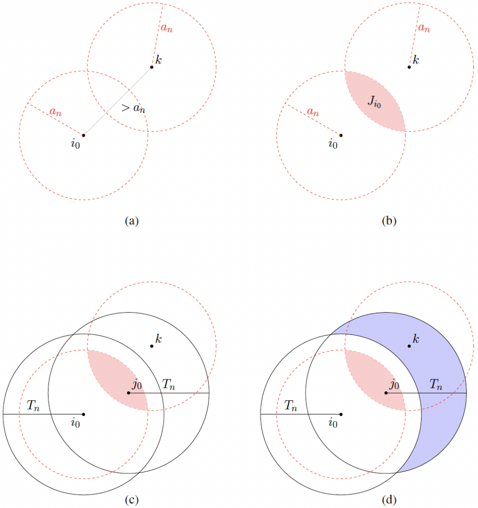


FIGURE 1.—Construction of the sensitivity argument. (a) Balls $B(i_0, a_n)$ and $B(k, a_n)$ for some node k at distance $> a_n$ from i_0. (b) The intersection J_{i_0} (shaded red). (c) For $j_0 \in J_{i_0}$, draw balls of radius T_n around i_0 and j_0. (d) The blue region is reachable from j_0 in T_n steps but not from i_0.

3. Forecasting Difficulties

Forecasting Setup

We assume i_0 and L_n are known perfectly.

The errors come from using the observed L_n as a stand-in (mistakenly assuming $E_n \equiv 0$),

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

A benchmark for $\hat{Y}_T(L_n)$ is using percolation on G_n and integrating over E_n rather than treating it as known

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

Forecasting Error

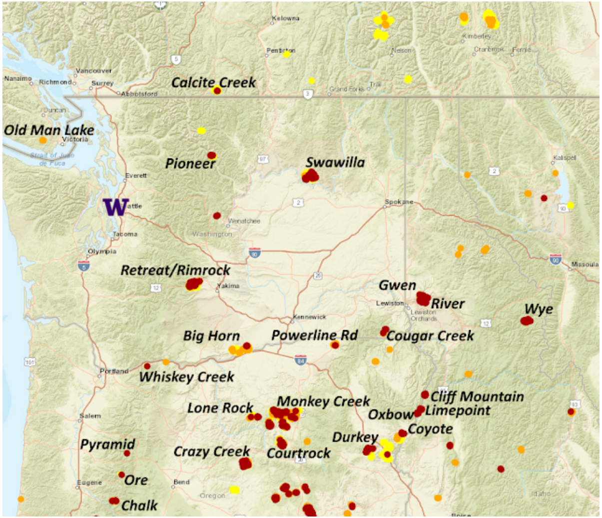
Theorem 2. (Extent of undercounting) Under Assumptions 1-3, as $n \to \infty$, $\frac{\hat{Y}_T(L_n)}{\tilde{Y}_T(G_n)} \to 0$.

Despite the advantages with perfect knowledge of not only L_n, i_0, T and q, the error will swamp the forecast as $n \to \infty$.

Small errors caused by the error network E_n recursively compound on themselves:

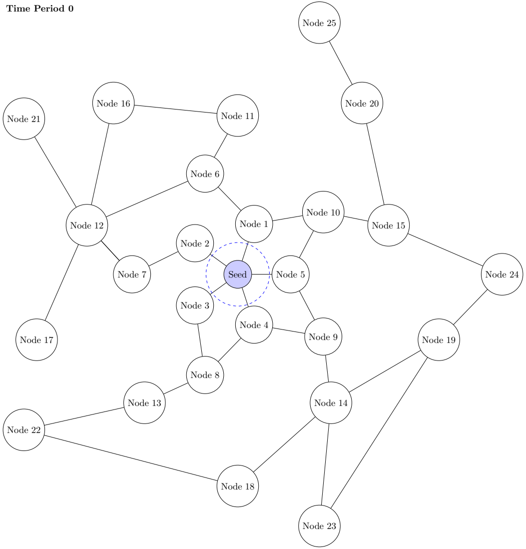
- As time grows, the volume around the seed grows in size, and the likelihood of hitting a mismeasured link in E_n increases.
- This leads to the creation of new activated regions elsewhere on the graph.
- In totality, these regions of activations caused by the propagating error dwarf the diffusion captured by the observed graph L_n.

Proof Sketch in Pictures

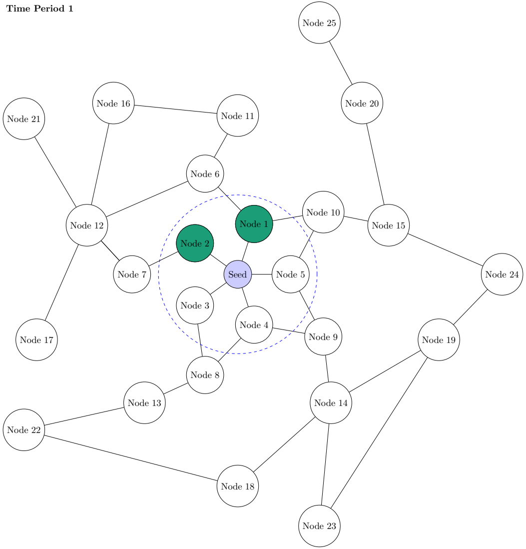


Source: WA state fire blog; 2024

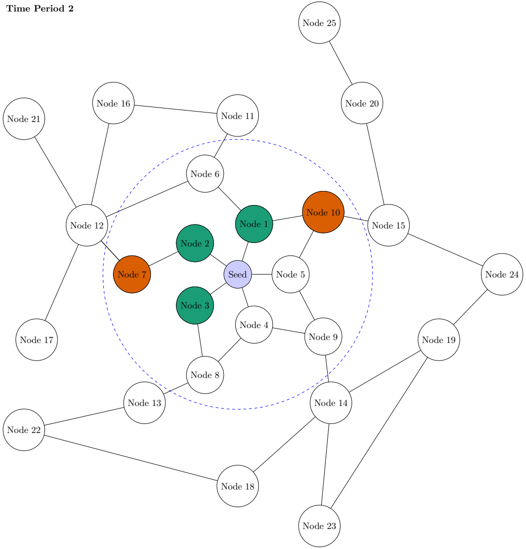
Proof Sketch in Pictures



Proof Sketch in Pictures

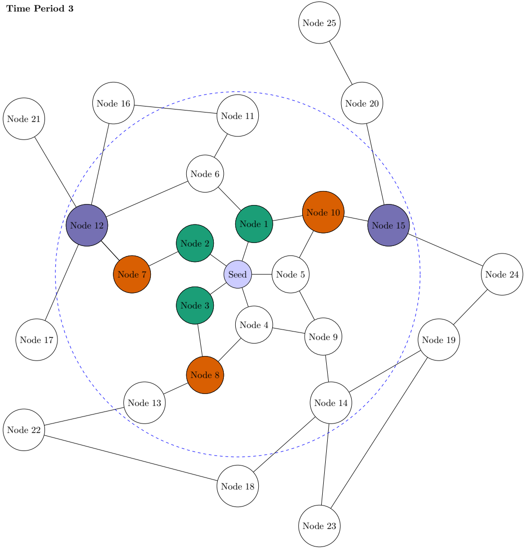


Proof Sketch in Pictures

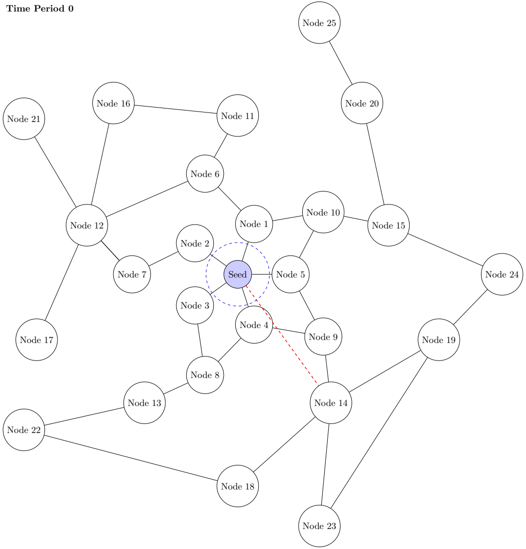


Proof Sketch in Pictures

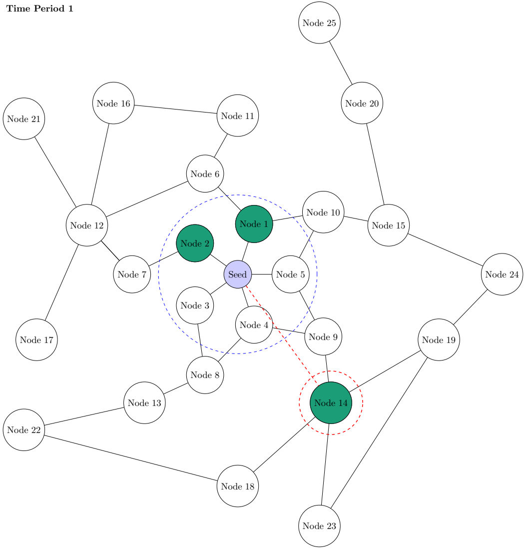
Time Period 3



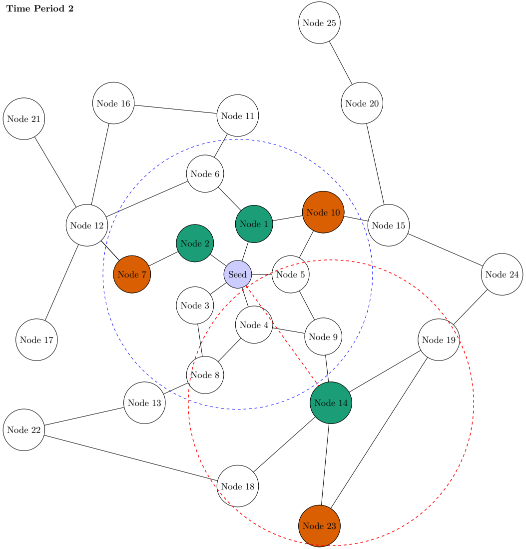
Proof Sketch in Pictures



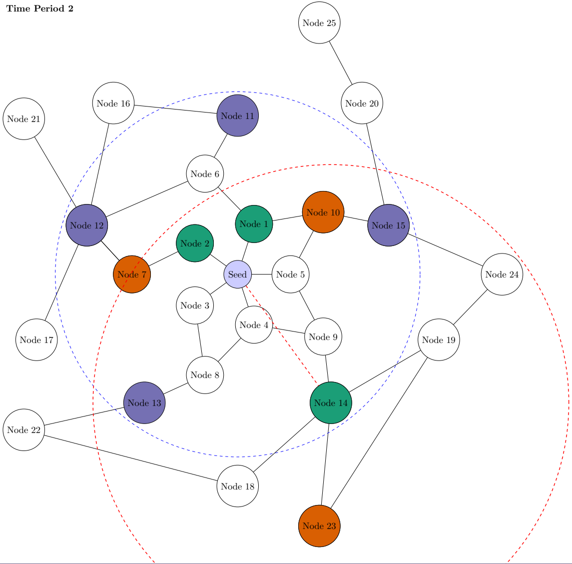
Proof Sketch in Pictures



Proof Sketch in Pictures



Proof Sketch in Pictures



4. Estimation and Possible Solutions

Estimating Parameters of the Process

Example: $\mathcal{R}_0$

- say $\hat{p}$ consistent for p_n and d_L (mean degree of L) known
- then $\hat{\mathcal{R}}_0 := \hat{p} d_L$ is consistent
- but still under Assumptions 1-3: sensitive dependence + forecasting problems

Bigger point: aggregative quantities (e.g., $\mathcal{R}_0, p_n$) may be easy to get

- still not enough for policy
- maybe good retrospective descriptives

Maybe we can "take better measurements"?

Possible Solution by Estimating β_n

Let's try to estimate β_n naively.

- Sample m_n nodes uniformly at random out of the n and perfectly observe $G_{ij,n}$.
- A sample of size m_n nodes will deliver $\binom{m_n}{2}$ possible links.
- In this way, links in E_n can potentially be observed to supplement the information of the known L_n.

But for our very small β_ns which cause problems, is this okay?

Failure of Estimating β_n

Proposition 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

   $\mathbb{P}(|\hat{\beta}_n/\beta_n - 1| < \epsilon) < c.$

San Jose, pop. approx 1 million

- 1000 surveyed
  - detects essentially no links in E_n
- 41,000 surveyed
  - volatile estimates

Can we be more clever?

- In the iid case, can use "phantom activation" to estimate β_n: activations with no observed activated neighbors
- But this only works in the iid case $\Rightarrow$ assumed away the problem of *where* the E_n links could potentially go

Widespread Testing

Another potential solution is the use of widespread testing:

- conduct random tests instantaneously and uniformly throughout the entire society of n nodes
- detect the activations with i.i.d. probability α_n
  - $\alpha_n \to 0$ with increasing n
  - realistic thought experiment: limited testing resources, etc.
- goal: forecast where in society activated agents reside at a given time period.

We show that the number of true regions that are activated at some time period will be grossly underestimated.

Failure of Widespread Testing

Theorem 3.

1. Detection prob. $\alpha_n \to 0$; Time $T < \alpha_n^{-1/(q+1)}$
2. $K_T^\star$ expected number of regions activated at T; $\hat{K}_T$ expected number w/ observed activated agent

As $n \to \infty$,

$$\frac{\hat{K}_T}{K_T^\star} \leq \underbrace{\alpha_n}_{\text{supply}\times\text{consent}\times\text{test power}} T^{q+1} < 1.$$

Ex.: Haryana, India – first 30 days

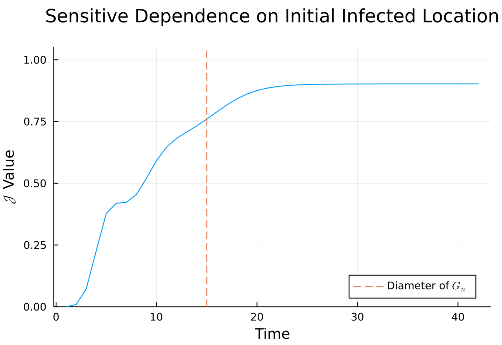
- Back of envelope calculation
  - conservative: maximum num. tests over first 3 months assumed to be done every day over the first month
  - actual policy: $\hat{K}_T / K_T^\star < 0.1$
- Counterfactuals
  - perfect power: $\hat{K}_T / K_T^\star < 0.15$
  - quintuple budget: $\hat{K}_T / K_T^\star < 0.75$

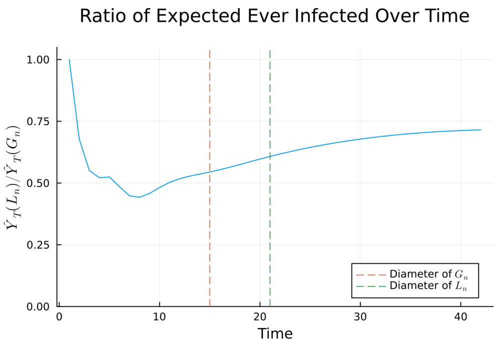
Govt. misses large share of regions with active agents over the first month

5. Empirical Applications

Mobility Flow Data for COVID-19 Pandemic

- Daily and weekly dynamic origin-to-destination population from (anonymized) mobile phone data (Kang et al., `20)
- Tract-to-tract flows starting March 1st, 2020 in the Southwest US (CA, NV, AZ).
- Construct L_n by linking tracts if the average flow between them (averaging over directions) is greater than 6 trips (the 93rd percentile of all flows).
- 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 (18% from the G_n^{92} graph).
- $\mathcal{J}(t)$ is a Jaccard index tracking the set of ever-activated nodes infected by an epidemic that begins from i_0 and j_0





- Alternate seeds within 2 links of i_0 (1.57% of pop).
- Elligible alternates, J_{i_0}: 82% of all within 2
- Error: cutoff at 92 vs 93 percentile of cross- census tract flows
- Get as bad as only estimating 48% of actual diffusion

Cai et al.: Informal Insurance in China

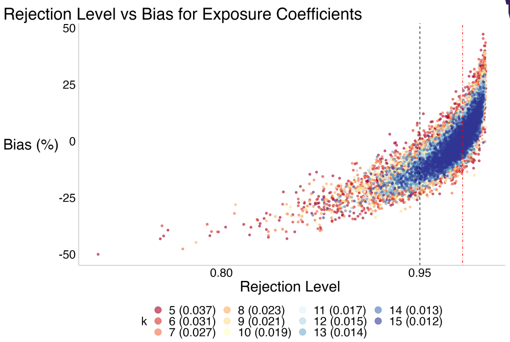
- Insurance products very important
- Seed info., generate a diffusion
- Outcome: take-up
- Core regressor: diffusion exposure

$$\text{diffusion exposure} = \left(\sum_{t=1}^{T} (p_n G)^t \right) s, \; s = \text{ seed indicator vector}$$

- take-up $= \beta$diffusion exposure $+$ stuff $+ \epsilon$
- $H_0 : \beta = 0$?
- look at tiny amounts of measurement error 1-4%
- note: in data, top-code at 5, avg degree 4.5

take-up $\sim$ diffusion exposure + stuff

	Insurance Uptake
Diffusion Exposure	0.029
	(0.012)
Household Controls	Yes
Village FE	Yes
Num Obs.	2676
Uptake Mean	0.459



- even with 1% error, bias has std. dev. of 8pp
- 3.7% bias, biases over 20% common
- fail to rej. H_0: no peer effect 15% of time!

6. Discussion

Discussion

- Small error in i_0 or G can cause major problems for:
  - where the diffusion goes? how much diffusion there is?
  - devising (practical) ambitious, localized policy solutions
- Contrast of aggregate vs. non-aggregate estimands:
  - Some aggregated quantities like $\mathcal{R}_0$ or p are still estimable
  - Local prediction is very hard
  - Suggests limited policy relevance for targeted prediction
- Implications beyond our diffusion setting...
  - lots of behavior has "percolation-like" foundations to exposure maps
    - coalition proof risk-sharing, public goods, p-common knowledge, referrals...
  - similar errors in exposure maps are almost guaranteed
    - how bad can they get?

Thank you!!