Non-Robustness of Diffusion Estimates on Networks with Measurement Error

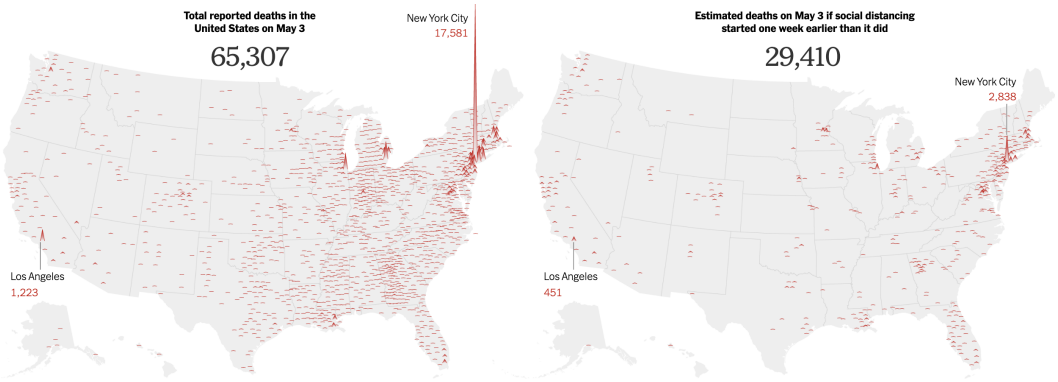
Arun Chandrasekhar, Stanford
Paul Goldsmith-Pinkham, Yale SOM
Tyler H. McCormick, University of Washington
Samuel Thau, Stanford
Jerry Wei, University of Washington

Northwestern Economics, 10/1/2024

Models of Diffusion on Networks

- Researchers and policymakers studying the spread of ideas, technology, or disease often estimate models of diffusion using network data
  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
     - where to seed new information to maximize spread
     - where to lockdown to prevent spread
  4. estimating counterfactuals
     e.g., estimates of peer effects.
- Goal today: show how tiny errors in network data can lead to large errors in diffusion models
- Natural example: SEIRD models during Covid-19

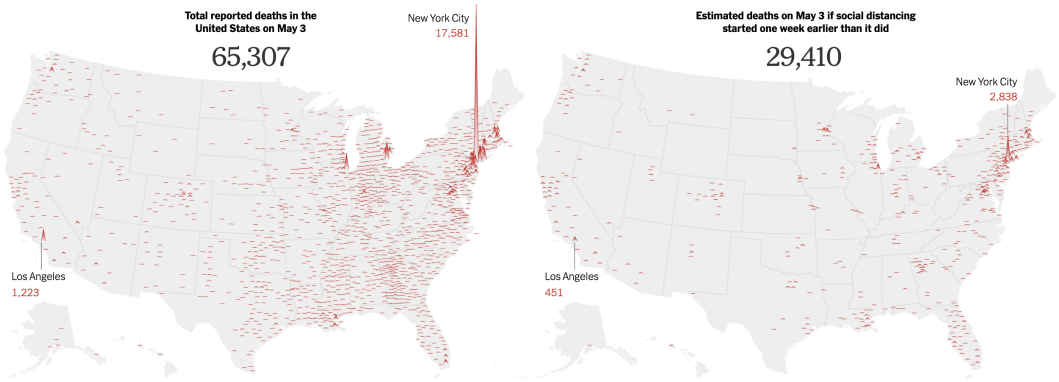
Counterfactual predictions of Covid-19 infections + deaths



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

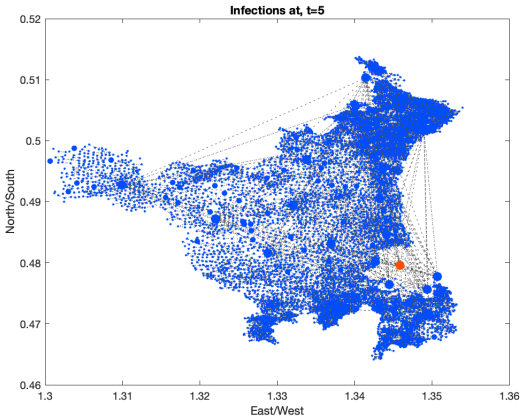


Modelers find that tens of thousands of U.S. deaths could have been prevented.

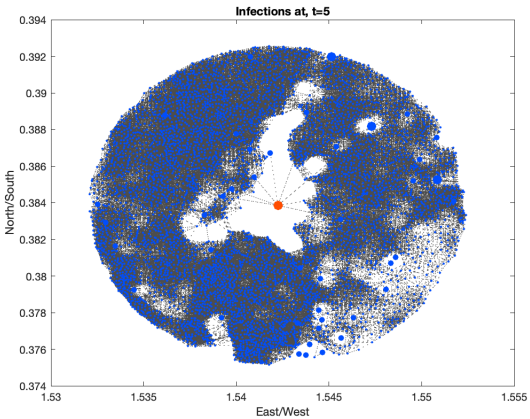
How did they model this?

Example: Rich Covid-19 SIERD Model in India

Haryana

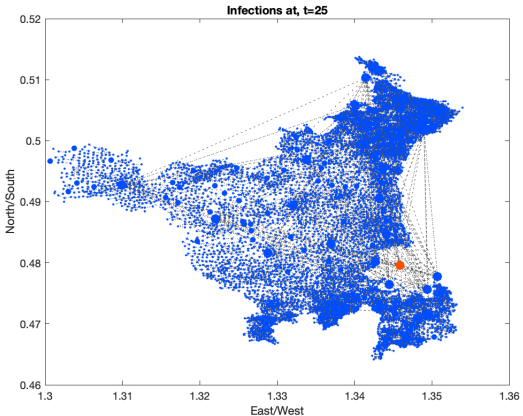


West Bengal (zoomed on Kolkata)

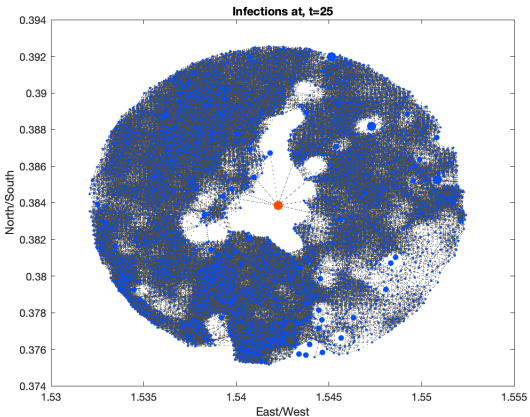


Example: Rich Covid-19 SIERD Model in India

Haryana

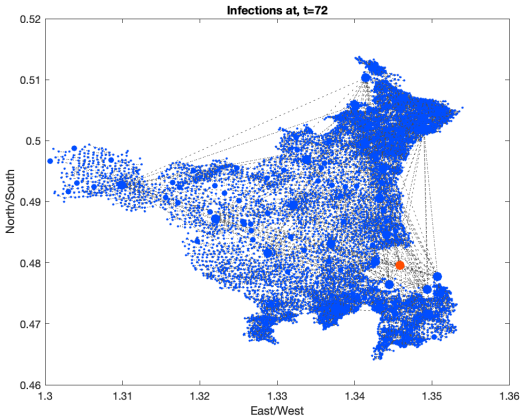


West Bengal (zoomed on Kolkata)

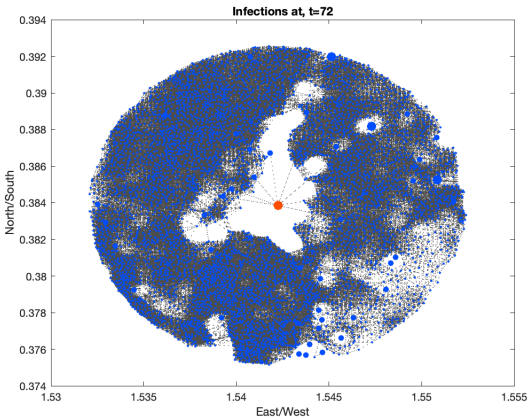


Example: Rich Covid-19 SIERD Model in India

Haryana

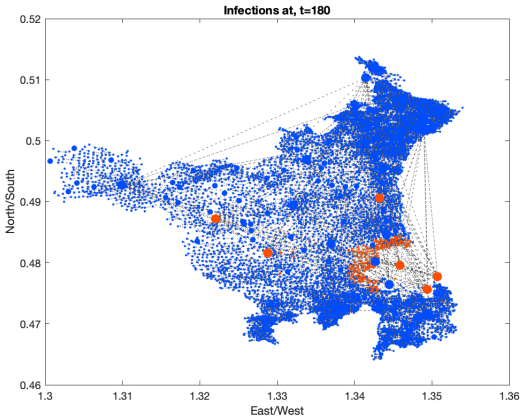


West Bengal (zoomed on Kolkata)

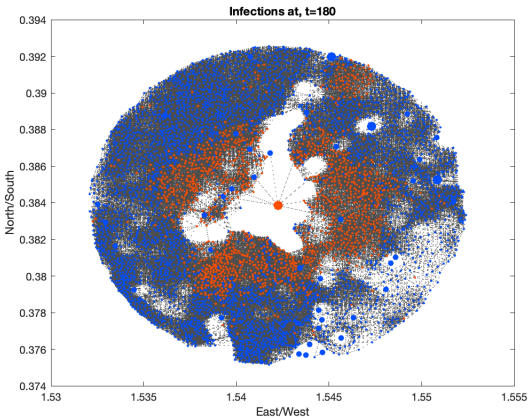


Example: Rich Covid-19 SIERD Model in India

Haryana

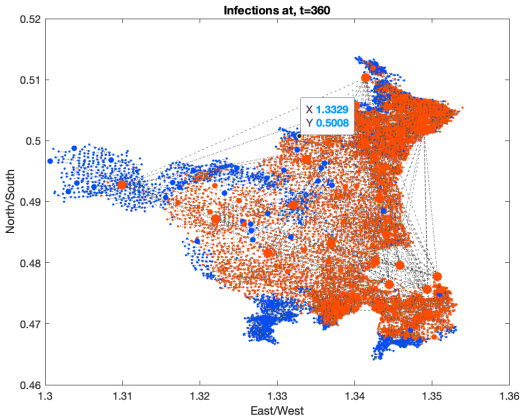


West Bengal (zoomed on Kolkata)

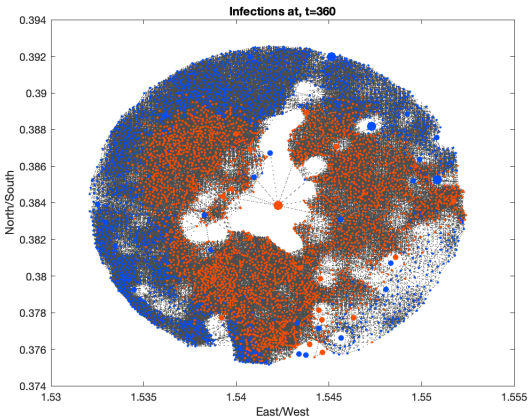


Example: Rich Covid-19 SIERD Model in India

Haryana

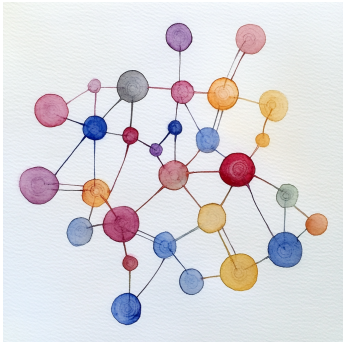


West Bengal (zoomed on Kolkata)



Two ingredients for diffusion on a network

Network connections

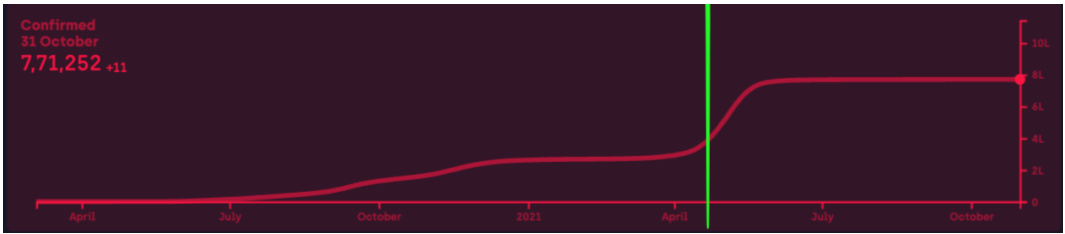


Diffusion process

- Quite plausible that the diffusion inputs or network connections are mismeasured
  - Who is patient zero?
  - Network connections measured with surveys or cell phone data

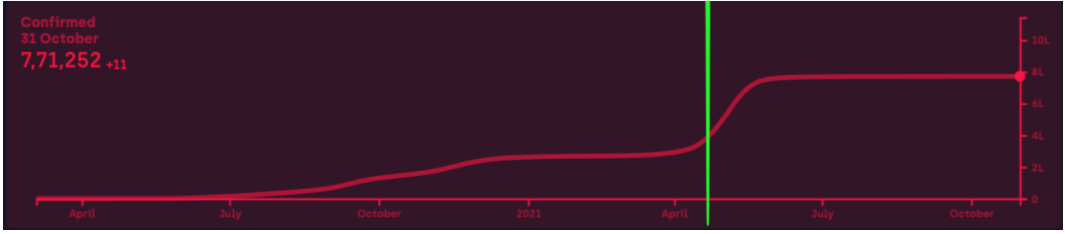
Forecasting a diffusion process in the medium run

- Goldilocks problem:
  - Day 1 is direct and uninteresting
  - In the long-run, diffuses everywhere
- scope for intervention is in the “sleeve” in the prior to the explosion



Forecasting a diffusion process in the medium run

- Goldilocks problem:
  - Day 1 is direct and uninteresting
  - In the long-run, diffuses everywhere
- scope for intervention is in the “sleeve” in the prior to the explosion



- Network structure and 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 these 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 these 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.

Percolation $P_n(G_n)$: directed graph, each link in G_n activated i.i.d. w/ prob. p_n.

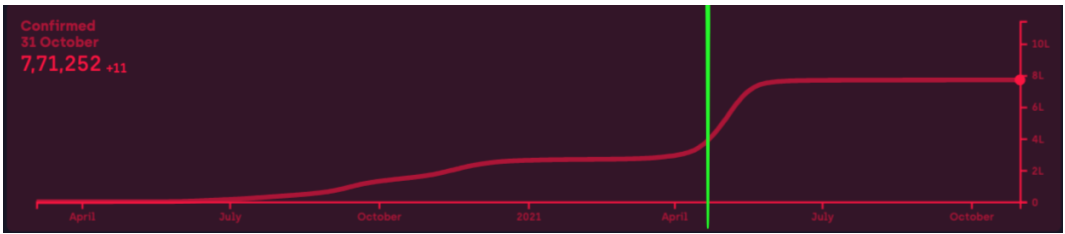
The Diffusion Process

The ever-activated set: $\mathcal{E}_t := \mathbb{E}\,|\{j \mid j \text{ ever activated by the diffusion on } L_n\}|$.
The shell: expected increment of new activations, $\mathcal{S}_t = \mathcal{E}_t - \mathcal{E}_{t-1}$.

Assumption 1. (Growth rates)
For some constant $q > 1$ and $t \in \mathbb{N}$,

1. $\mathcal{E}_t = \Theta(t^{q+1})$ and $\mathcal{S}_t = \Theta(t^q)$.
2. $p_n \in \left((\log n)^{-q/(2q+2)}, 1\right]$

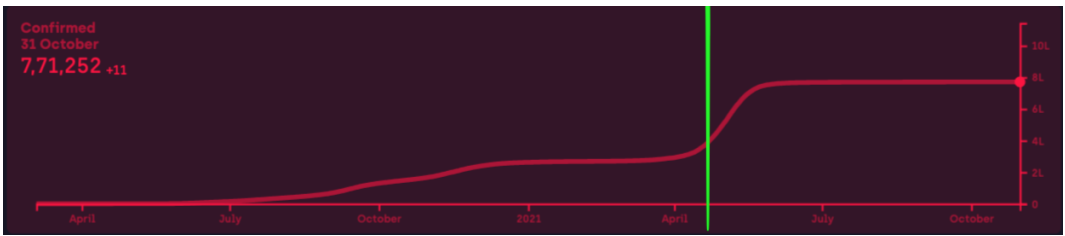
1. $\mathcal{E}_t$: polynomial growth period
2. p_n: some diffusion happens



Policy Relevant Time Period

Assumption 2. (Time Period of Focus)
Let a be any positive constant satisfying $2a > 1/(q+1)$.
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) $\underline{T}_n = (\log n)^a$.

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...
   - more expansive ($q \uparrow$), the earlier the party gets started



Missing Mechanism

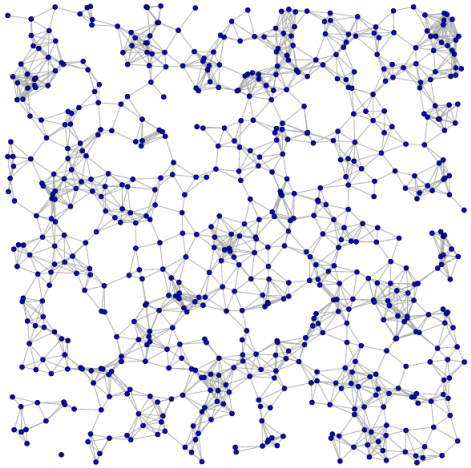
A given node i can link to fraction $\delta_n \in (\underline{\delta}_n, 1]$. of nodes through E_n.

Assumption 3. (Missing Mechanism) Let $2a > 1/(q+1)$ and $\nu := a - 1/(2q+2)$. The lower bound on the share of nodes that can be linked to is given by

$$\underline{\delta}_n = (\log n)^{-q\nu}.$$

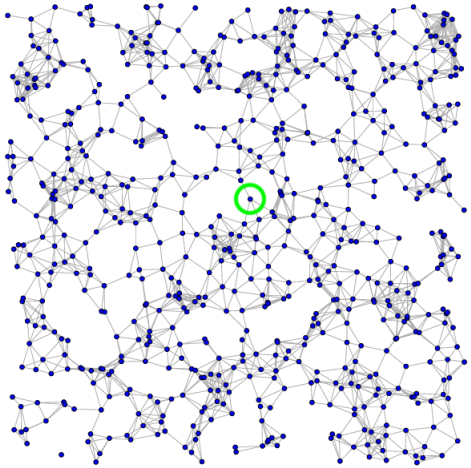
1. Take somewhat flexible stand as to which js constitute this δ_n-share.
2. May be that the entire δ_n share of nodes are in a highly localized neighborhood (in L_n) of i.
3. Results not about long-range "shortcuts."

Missing Mechanism



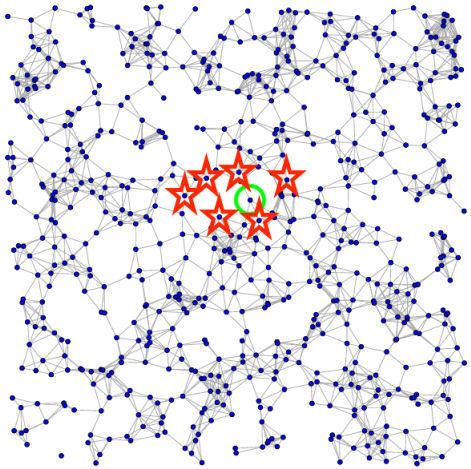
stylized L_n

Missing Mechanism



seed i_0

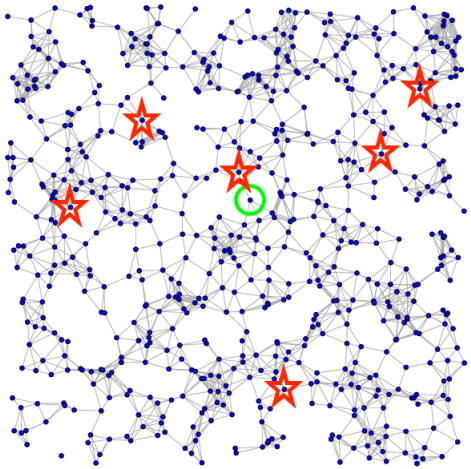
Missing Mechanism



local support:

- i_0 can only make links in E_n with ★
- local in L_n

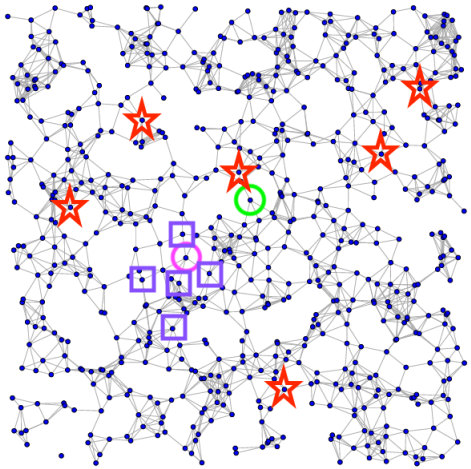
Missing Mechanism



global support:

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

Missing Mechanism



two seeds i_0 and j_0:

- 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
- lower bound close to 1 in either case
- $\underline{\delta}_n < 0.001$ and even lower ok
- structures w/ rare links, local in L_n,

Haryana: pop 25.4 million

- $q = 2$: upper bound 10 months
- $q = 3$: 2.4 months
- lower bound close to 1 in either case
- $\underline{\delta}_n < 0.001$ and even lower ok
- structures w/ rare links, local in L_n,

Missing Probability

Assumption 4. (Missing Probability) For every n, i, j, $E_{ij} \sim \mathrm{Ber}(\beta_n)$ for up to some share δ_n of the n nodes and is zero otherwise. Further,

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

- live in interesting case: $\beta_n = o(n^{-1})$ means E_n not connected / no giant comp.

2. Sensitive Dependence to Seed Set

known G, locally perturbed $i_0 \to j_0$

Notation for keeping track of activations

- $B^G_{i_0}(T_n)$: set of nodes j that could have been activated by T_n is the ball in G_n (sometimes call these “catchment regions”)

Notation for keeping track of activations

- $B^P_{i_0}(T_n) \subset B^G_{i_0}(T_n)$: set of nodes j that are activated by T_n is the ball in P_n

Notation for keeping track of activations

- $B^P_{i_0}(T_n) \subset B^G_{i_0}(T_n)$: set of nodes j that are activated by T_n is the ball in P_n
- local neighborhood $U_{a_n}(i_0) \subset B^G_{i_0}(T_n)$: a ball of radius $a_n/T_n \to 0$
- set of j

$$J_{i_0} := \left\{ j : \exists k \notin B^G_{i_0}(T) \text{ with } U_{a_n}(i_0) \cap U_{b_n}(k) \right\}$$

  (a) local to i_0;
  (b) local to some other k;
  (c) k can't be reached by i_0 in T periods

- Compare $B^P_{i_0}(T_n)$ to $B^P_{j_0}(T_n)$ and show they can differ considerably

$$\Delta(i_0, j_0) := |B^P_{i_0}(T_n) \cap B^P_{j_0}(T_n)| / |B^P_{i_0}(T_n) \cup B^P_{j_0}(T_n)|, \text{ Jaccard Index}$$

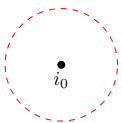
Sensitive Dependence

Theorem 1. Let Assumptions 1-3 hold and i_0 be a seed. The local neighborhood vanishes– $|U_{a_n}(i_0)|/n \to 0$ and wpa1 (over (P_n, E_n))

1. a non-vanishing share belongs to J_{i_0}: $|J_{i_0}|/|U_{a_n}(i_0)| > c$
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



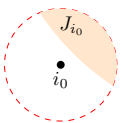
Sensitive Dependence

Theorem 1. Let Assumptions 1-3 hold and i_0 be a seed. The local neighborhood vanishes– $|U_{a_n}(i_0)|/n \to 0$ and wpa1 (over (P_n, E_n))

1. a non-vanishing share belongs to J_{i_0}: $|J_{i_0}|/|U_{a_n}(i_0)| > c$
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



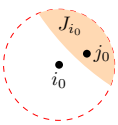
Sensitive Dependence

Theorem 1. Let Assumptions 1-3 hold and i_0 be a seed. The local neighborhood vanishes– $|U_{a_n}(i_0)|/n \to 0$ and wpa1 (over (P_n, E_n))

1. a non-vanishing share belongs to J_{i_0}: $|J_{i_0}|/|U_{a_n}(i_0)| > c$
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



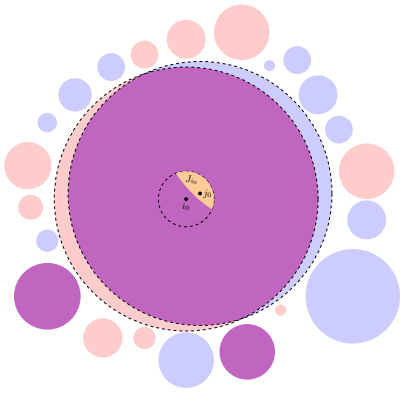
Sensitive Dependence

Theorem 1. Let Assumptions 1-3 hold and i_0 be a seed. The local neighborhood vanishes– $|U_{a_n}(i_0)|/n \to 0$ and wpa1 (over (P_n, E_n))

1. a non-vanishing share belongs to J_{i_0}: $|J_{i_0}|/|U_{a_n}(i_0)| > c$
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



3. Forecasting Difficulties

imperfect knowledge of G, known i_0

Forecasting Setup

- Assume i_0 and L_n are known perfectly.
- Econometrician mistakenly uses the observed L_n instead of G_n:

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

- Benchmark is targeting G_n (integrating over E_n error):

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

Forecasting Error

Theorem 2. (Extent of undercounting)
Under Assumptions 1-4, as $n \to \infty$, $\frac{\hat{Y}_T(L_n)}{\hat{Y}_T(G_n)} \to 0$.

- Despite perfect knowledge of L_n, i_0, T and q, the error dominates as $n \to \infty$.
- Why? Small errors caused by E_n recursively compound on themselves.
- As time grows, the volume around the seed grows.
- Then, the likelihood of hitting a link in E_n increases.

Forecasting Error

Theorem 2. (Extent of undercounting)
Under Assumptions 1-4, as $n \to \infty$, $\frac{\hat{Y}_T(L_n)}{\hat{Y}_T(G_n)} \to 0$.

- Despite perfect knowledge of L_n, i_0, T and q, the error dominates as $n \to \infty$.
- Why? Small errors caused by E_n recursively compound on themselves.
- As time grows, the volume around the seed grows.
- Then, the likelihood of hitting a link in E_n increases.
- This creates new activated regions elsewhere on the graph.
- These new regions created by jumps repeat the same process.
- The error caused by these new region swamp the forecast process.

3b. Forecasting in the Exponential Case

imperfect knowledge of G, known i_0

New assumptions for the exponential regime

Assumption 1→5

For some constant $q > 1$ and $t \in \mathbb{N}$,

1. $\mathcal{E}_t = \Theta(q^t)$ and $\mathcal{S}_t = \Theta(q^t)$.
2. $p_n \delta_n > \frac{1}{\log(n)}$

1. $\mathcal{E}_t$: Exponential growth period
2. p_n: much lower passing rate allowed

New assumptions for the exponential regime

Assumption 1→5

For some constant $q > 1$ and $t \in \mathbb{N}$,

1. $\mathcal{E}_t = \Theta(q^t)$ and $\mathcal{S}_t = \Theta(q^t)$.
2. $p_n \delta_n > \frac{1}{\log(n)}$

Assumption 2 → 6

$T_n \in [\underline{T}_n, \overline{T}_n]$ where:

1. $\overline{T}_n = \log(n)$ and
2. $\underline{T}_n = (\log \log n)$.

1. $\mathcal{E}_t$: Exponential growth period
2. p_n: much lower passing rate allowed

1. $\underline{T}_n$: $\log \log n$ is much earlier
2. $\overline{T}_n$: $n^{\frac{1}{q+1}}$ changed to $\log(n)$

New assumptions for the exponential regime

Assumption 1→5

For some constant $q > 1$ and $t \in \mathbb{N}$,

1. $\mathcal{E}_t = \Theta(q^t)$ and $\mathcal{S}_t = \Theta(q^t)$.
2. $p_n \delta_n > \frac{1}{\log(n)}$

1. $\mathcal{E}_t$: Exponential growth period
2. p_n: much lower passing rate allowed

Assumption 2 → 6

$T_n \in [\underline{T}_n, \overline{T}_n]$ where:

1. $\overline{T}_n = \log(n)$ and
2. $\underline{T}_n = (\log \log n)$.

1. $\underline{T}_n$: $\log \log n$ is much earlier
2. $\overline{T}_n$: $n^{\frac{1}{q+1}}$ changed to $\log(n)$

Assumption 4 → 7

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

1. $\delta_n \in (\underline{\delta}_n, 1]$
2. $\beta_n = \Omega(\frac{1}{p_n \delta_n n})$

1. If $p_n \delta_n < 1$, then E_n will have a giant component

Forecasting Error

Theorem 2b. (Extent of undercounting)
Under Assumptions 4, 5-7, as $n \to \infty$, $\frac{\hat{Y}_T(L_n)}{\hat{Y}_T(G_n)} \to 0$.

- The same logic holds here, but now we need more idiosyncratic links
- Due to dominant explosion from expontential growth, need errors to grow fast
- In fact, giant component in E_n is crucial

Forecasting Error

Theorem 2b. (Extent of undercounting)
Under Assumptions 4, 5-7, as $n \to \infty$, $\frac{\hat{Y}_T(L_n)}{\tilde{Y}_T(G_n)} \to 0$.

- The same logic holds here, but now we need more idiosyncratic links
- Due to dominant explosion from expontential growth, need errors to grow fast
- In fact, giant component in E_n is crucial

Proposition. (Partial converse) Assume that L_n is made up of K_n independent regions, which each satisfy Assumption 5. Then, if $\beta_n = O(\frac{1}{p_n \delta_n n})$, we have that $\frac{\hat{Y}_T(L_n)}{\tilde{Y}_T(G_n)} \to 1$.

4. Estimation and Possible Solutions

aggregate quantities often ok; sampling/testing not

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 1-4: sensitive dependence + forecasting problems

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

- still not enough for targeted 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. Let $\delta_n = 1$ (i.i.d. random error in E_n).

- 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.
- New info in E_n can supplement the information of the known L_n.

Does this solve the problem for our our very small β_n?

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?

- perhaps... but the problem is also harder
- recall: assumed away the problem of *where* the E_n links could potentially go

Widespread Testing

Another potential solution is the use of widespread testing:

- goal: forecast where the activation (disease) has gotten as of period T
  - Consider the network split into disjoint regions (e.g. states or cities)
  - Where has it gotten?
- Instantaneously perform random samples throughout the n nodes
- Detect the activations with i.i.d. probability α_n
  - $\alpha_n \to 0$ with increasing n
  - thought experiment: combination of limited supply, testing consent, and test power

$\to$ Number of true regions that are activated at some time period will be grossly underestimated.

Failure of Widespread Testing

Theorem 3.

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

- Govt. very proactive
- 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: $\hat{K}_T / K_T^\star < 0.75$

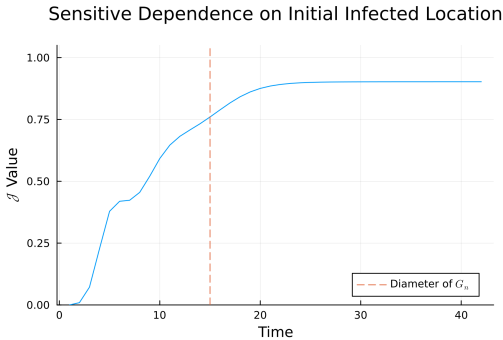
Govt. misses large share of regions with active agents over the first month

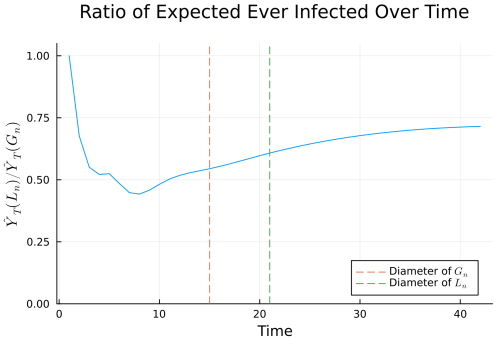
5. Empirical Applications

California Covid-19, ~~India marketing~~, China insurance

Mobility Flow Data for COVID-19 Pandemic

- Mobile phone data mapping origin-to-destination flows
  - anonymized data (Kang et al., '20)
- We use tract-to-tract flows from March 1st, 2020 in the Southwest US (CA, NV, AZ).
  - Network is extremely dense
- Construct L_n across census tracts
  - by linking tracts if the average flow between them (averaging over directions) is greater than six 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





- U_{i_0}: 1.57% of pop.
- J_{i_0}: 82% of U_{i_0}
- $\mathcal{J}$ ($\approx \Delta(i_0, j_0)$): agreement quite low!

- Error: cutoff at 92 vs 93 percentile of cross- census tract flows
- Get as bad as only estimating 48% of actual diffusion

More provocative: Cai et al.: Informal Insurance in China

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

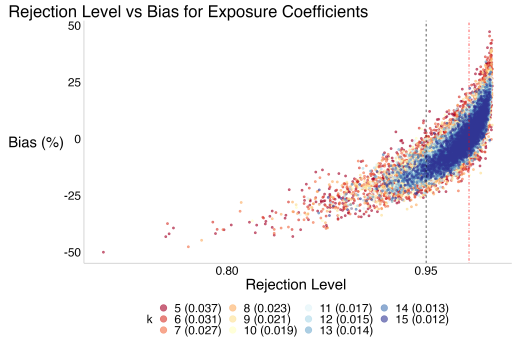
$$\text{diffusion exposure} = \left(\sum_{t=1}^{T} (p_n G)^t \right) s, \; s = \text{ seed indicator vector}$$

- 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

$p = 0.02$ in the data



- 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
    - difficult for “welfare analysis” exercises (not this talk)
- 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?
    - Perhaps a question of rate of diffusion vs. rate of error