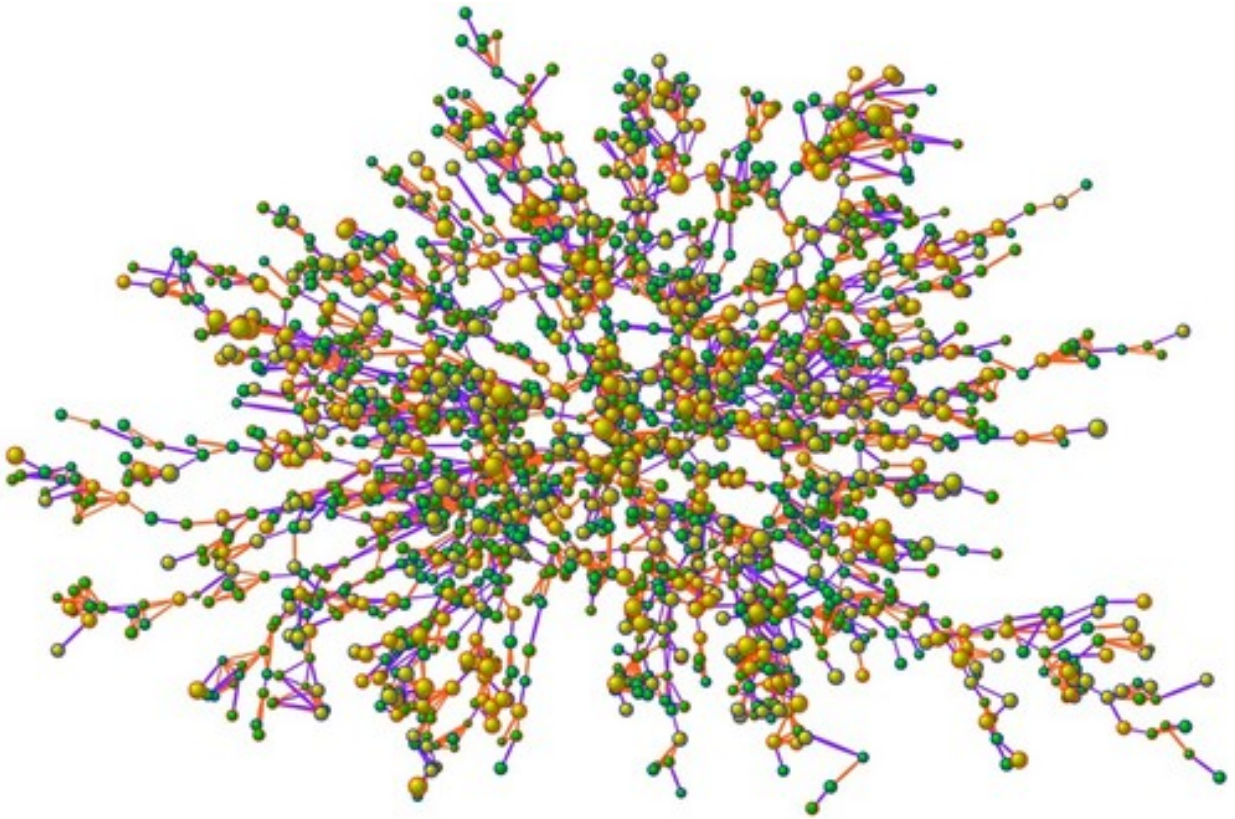
Causal and statistical inference in the presence of network dependence

Elizabeth L. Ogburn

Department of Biostatistics,
Johns Hopkins University

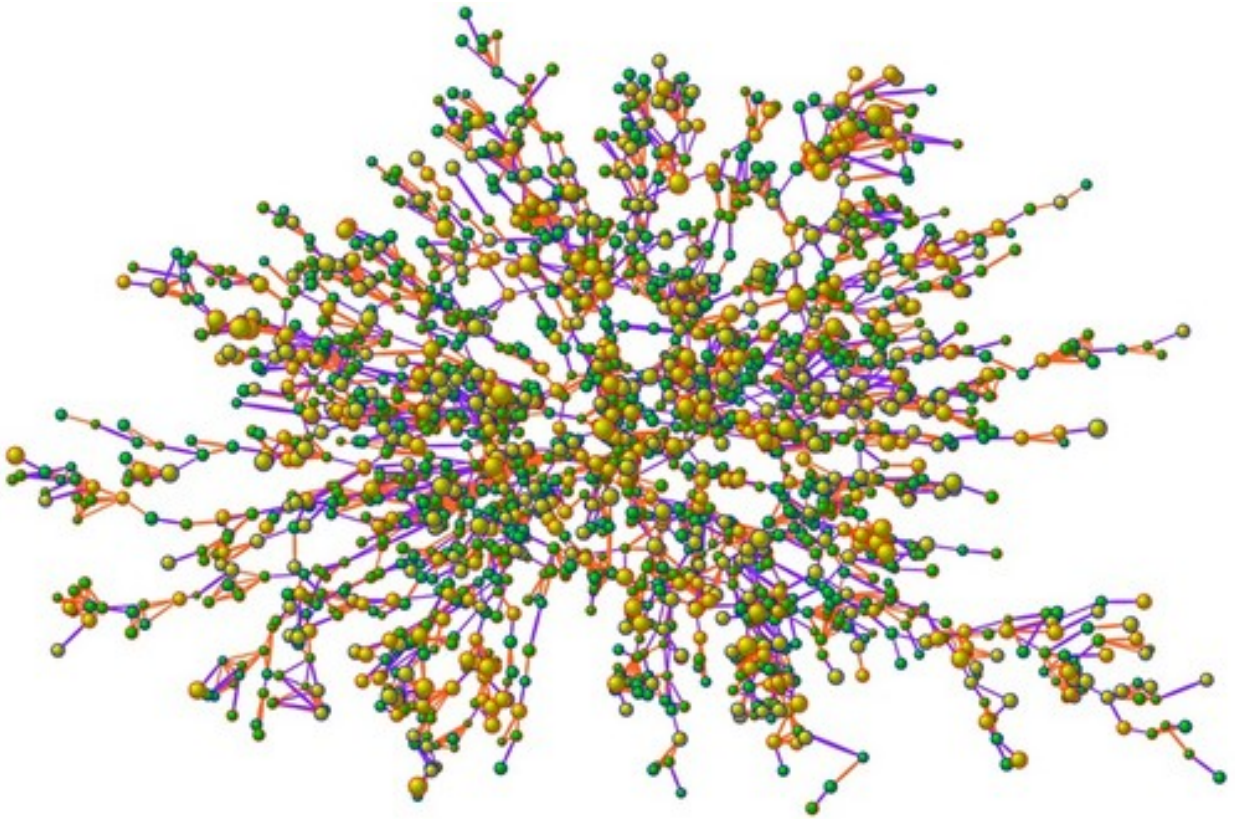
outline

- Brief history of causal inference using network data.
- What is network dependence, and why is it a problem?
- Two partial solutions:
  1. Subsample conditionally independent observations
     - naive but easy to understand, implement, and generalize,
     - dependence due to contagion;
  2. Semiparametric approach based on the efficient influence function
     - more sophisticated and powerful but less intuitive and difficult to implement.



www.nicholaschristakis.net

- Each node is associated with an outcome, treatment, covariates.
- Causal effects of interest include peer effects, treatment effects, spillover/interference effects, effects of network interventions, ...



www.nicholaschristakis.net

Two challenges for causal inference using network data:

- nonparametric identification of causal effects (interference, confounding by homophily, positivity violations),
- statistical inference in the presence of network dependence.

brief history of causal inference using network data

- Christakis and Fowler (2007, 2008, 2009, 2010, 2011, 2012) initiated a wave of interest in estimating peer effects from social network data.
  - To examine peer effects, they fit models

$$Y_{ego}^{t} \sim Y_{alter}^{t-1}, Y_{alter}^{t-2}, Y_{ego}^{t-2}, \boldsymbol{C}_{ego}$$

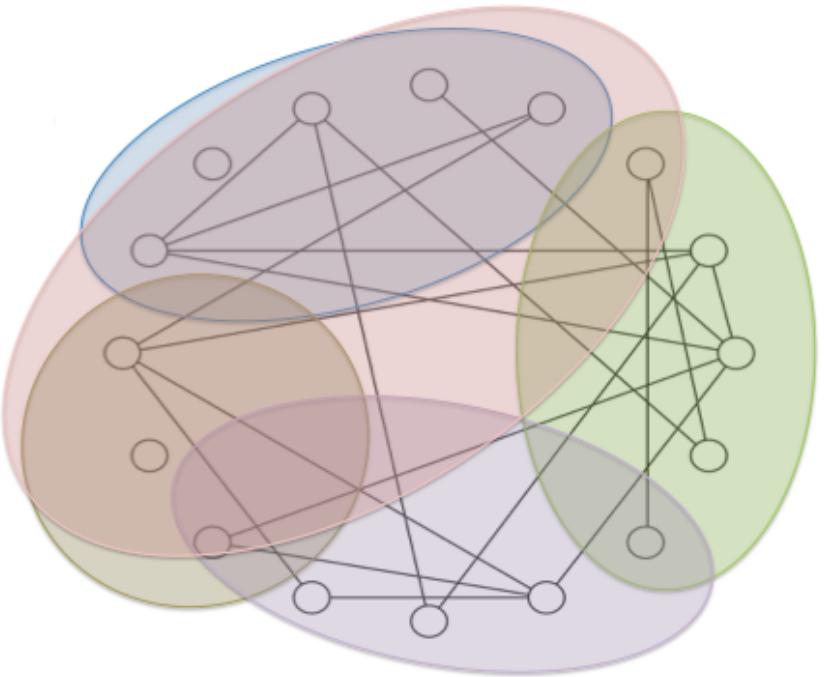
  - Widely publicized results include significant peer effects for obesity, smoking, alcohol consumption, sleep habits, etc.
  - Researchers began using similar models to assess peer effects across a wide range of disciplines and problems (e.g. Ali and Dwyer, 2009; Cacioppo et al., 2009; 2008; Lazer et al., 2010; Rosenquist et al., 2010, Wasserman 2012).

brief history causal inference using network data

- There is growing interest in randomization-based inference for networks (e.g. Toulis & Kao, 2013; Bowers et al., 2013; Aronow & Samii, 2013; Eckles et al., 2014, Choi 2016).
- Work on interference usually relies on randomization and on the assumption of **partial interference**, but may provide a solution to the problem of network dependence in cluster randomized trials (e.g. Sobel, 2006; Hong & Raudenbush, 2006; Rosenbaum, 2007; Hudgens & Halloran, 2008; Tchetgen Tchetgen & VanderWeele, 2012; Liu & Hudgens, 2014).
- Mathematical modeling of contagious processes avoids these problems but is highly dependent on parametric assumptions about agent-based processes (e.g. Steglich, Snijders & Pearson, 2007; Railsback & Grimm, 2011).

sources of network dependence

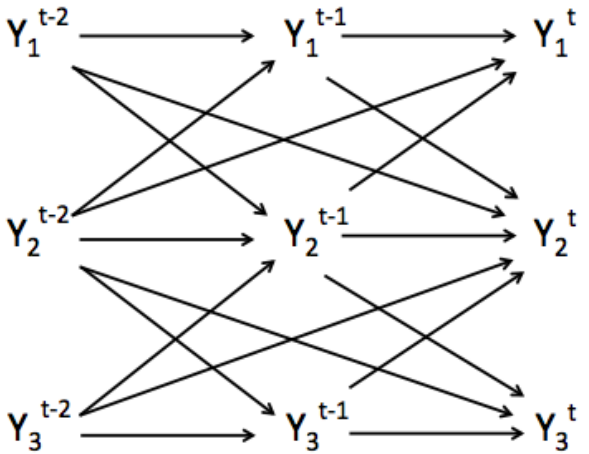
- **Latent variables** cause outcomes among close social contacts to be more correlated than among distant contacts. (E.g. homophily, geography, shared culture, shared genetics.)



- Similar to spatial dependence.

sources of network dependence

- **Contagion** implies information barrier structures, e.g. $\left[Y_1^t \perp Y_2^t \mid Y_1^{t-2}, Y_2^{t-2}, Y_1^{t-1}, \text{ and } Y_2^{t-1}\right]$ and $\left[Y_1^{t-2} \perp Y_3^{t-1}\right]$.



- When a network is observed at a single time point, this will resemble latent variable dependence.
- If the network is observed frequently, so that the outcome can't diffuse very far between observations, we can harness conditional independence restrictions to facilitate inference.

why is dependence a problem?

- Statistical analysis that incorrectly assumes independence will be invalid.
- This is a very hard problem when dependence is due to latent variables and is unstructured.
- It's not quite as hard when dependence is due to contagion.
- Two problems for traditional frequentist inference:
  - CLT may not hold,
  - Standard error estimates and resulting inference will be anticonservative.

- If $\bar{Y} \to \mu$, the rate of convergence is determined by

$$var(\bar{Y}) = \frac{1}{n^2}\left\{\sum_{i=1}^{n}\sigma^2 + \sum_{i\neq j} cov(Y_i, Y_j)\right\}$$

- Define

$$b_n = \frac{1}{n}\sum_{i\neq j} cov(Y_i, Y_j)$$

- Now

$$var(\bar{Y}) = \frac{\sigma^2}{n/\left(1+\frac{b_n}{\sigma^2}\right)}$$

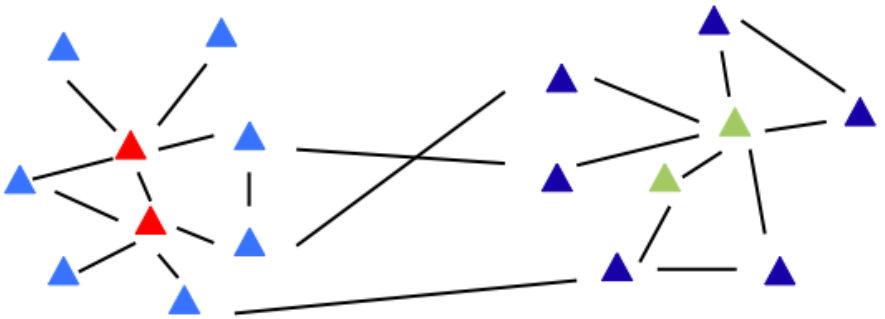
- If a CLT holds, then

$$\sqrt{\frac{n}{1+\frac{b_n}{\sigma^2}}}\left\{\bar{Y}-\mu\right\} \xrightarrow{d} N(0,\sigma^2)$$

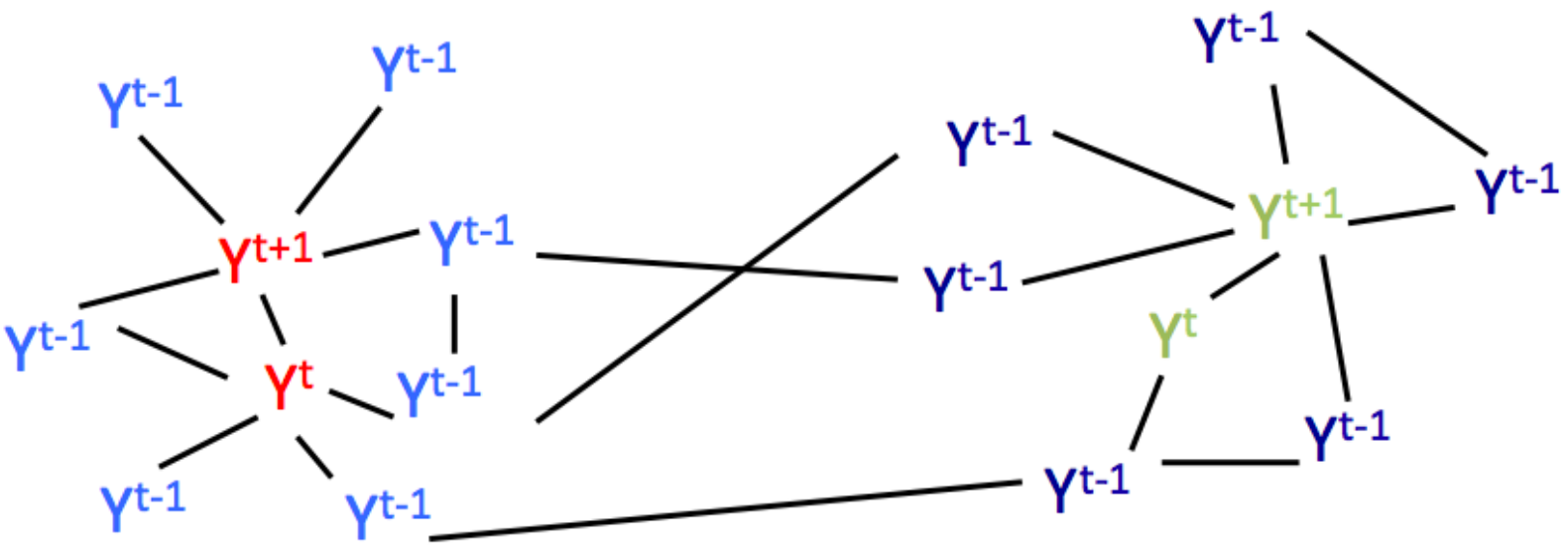
naive solution

joint work with Tyler VanderWeele

- Summary: Create conditionally independent units; analyze with standard, i.i.d. models, but **conditional** on "information barriers."
- Randomly sample non-overlapping groups from the network.



- This will allow us to condition on an "information barrier."
- Now we can estimate conditional estimands using standard statistical machinery like GLMs.
  - The residuals will be uncorrelated across subjects despite the dependence structure.



- Regress Y^{t+1} on Y^t conditional on $\{Y^{t-1}\}$

For details see Ogburn & VanderWeele, Vaccines, contagion, and social networks (forthcoming in AoAS)

naive solution

- Pros
  - easy to understand, easy to implement
  - generalizable to many estimands and models (in principle)
  - may be feasible if full data structure is unavailable, as long as information barriers can be found
- Cons
  - dependence due to contagion only
  - sample size $<$ true effective sample size
    - requires throwing away data
    - low power
  - estimand must be conditional
    - more appropriate for causal effects than for sample means

more principled solution

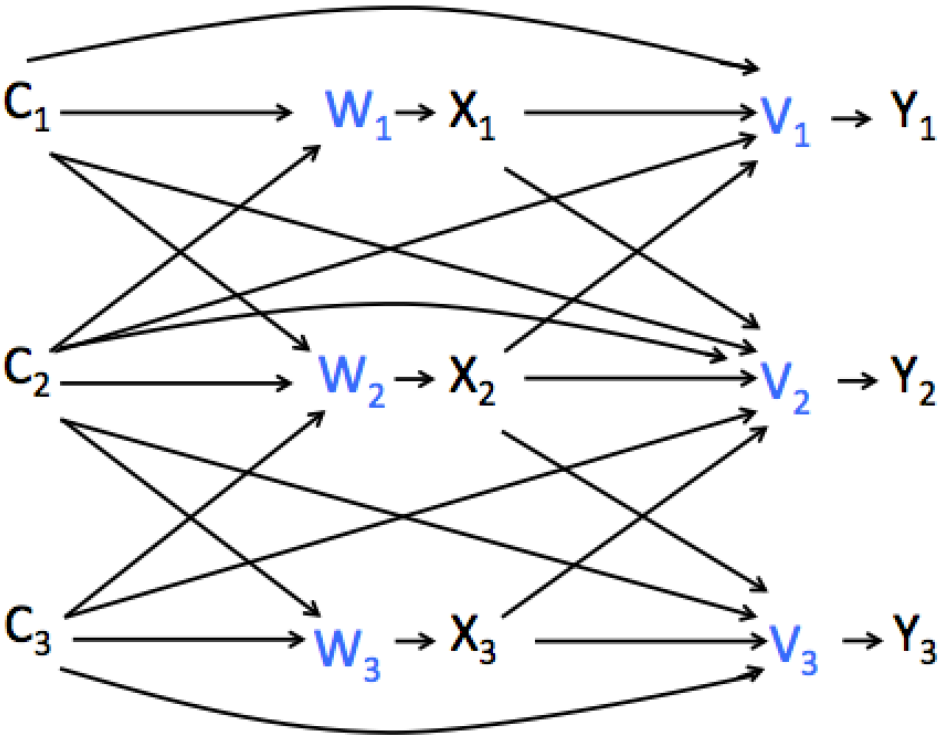
joint work with Oleg Sofrygin, Ivan Diaz, Mark van der Laan

- Extension of semiparametric, influence-function-based inference from the iid setting.
- We define a model $\mathscr{M}$, which restricts the observed data distribution in some way(s).
- We are interested in estimating a parameter ψ under model $\mathscr{M}$, i.e. a functional of the observed data.
- Under $\mathscr{M}$, there is a class of **influence functions** for ψ.
  - Each (RAL) estimator $\hat{\psi}$ is paired with an IF φ, and in the iid setting

$$\sqrt{n}(\hat{\psi} - \psi) \overset{p}{\approx} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(O_i)$$

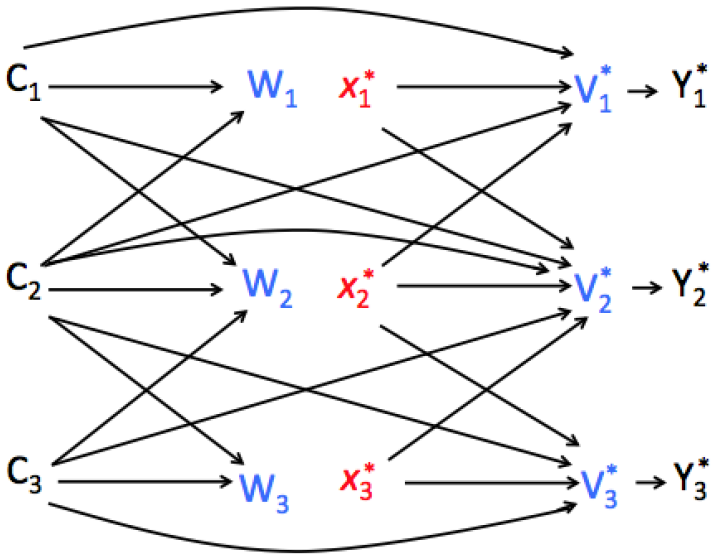
  - Because the IF has mean 0 at the true parameter value, we can use it to create unbiased estimating functions for ψ.

- van der Laan (2014) extended this approach to settings with interference and/or contagion.
  - Not partial interference, but each subject can only interfere with $\leq K$ other subjects.
- We extend van der Laan (2014) to social network settings:
  - K grows with n
  - highly connected "hubs" may exert undo influence
  - estimation of causal effects of interventions on features of network topology
- This framework can handle longitudinal data, but for simplicity we focus on the single-time-point setting.



- We make independence assumptions that entail
  - there is no unmeasured confounding,
  - $C_i \perp C_j$ if i and j have no friends in common,
  - $Y_i \perp Y_j | parents$ and $X_i \perp X_j | parents$ if i and j have no friends in common.

- The simplest kind of intervention deterministically sets $\mathbf{X}$ to a user-specified value $\mathbf{x}^*$:



- Y_i^* is the counterfactual outcome of individual i in a hypothetical world in which $P(\mathbf{X} = \mathbf{x}^*) = 1$.
  - Peer effects: X_i could be a function of alters' outcomes at a previous time point.
- We are interested in $E\left[\bar{Y}^*\right]$, where $\bar{Y}^* = \frac{1}{n}\sum_{i=1}^{n} Y_i^*$.

- $E\left[\bar{Y}^*\right]$ is identified by the parameter

$$\psi = \frac{1}{n}\sum_{i=1}^{n} E\left[\sum_y y\, p_Y(y|V_i^*)\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_v \left[\sum_y y\, p_Y(y|v)\right] P[V_i^* = v].$$

- The efficient influence function for ψ (in a particular semiparametric model) is

$$\varphi(\mathbf{O}) = \sum_{j=1}^{n}\frac{1}{n}\sum_{i=1}^{n} E\left[\sum_y y\, p_Y(y|V_i^*) \mid C_j = c_j\right] - \psi$$
$$+ \frac{1}{n}\sum_{i=1}^{n} \frac{\frac{1}{n}\sum_{j=1}^{n} P(V_j^* = v_i)}{\frac{1}{n}\sum_{j=1}^{n} P(V_j = v_i)} \left\{ y_i - \sum_y y\, p_Y(y|v_i)\right\}$$

- This is of the form $\frac{1}{n}\sum_{i=1}^{n}\varphi_i(\mathbf{O})$ instead of $\frac{1}{n}\sum_{i=1}^{n}\varphi(O_i)$.

- Turning the efficient IF into an estimating equation and solving it gives us an estimate $\tilde{\psi}$ of ψ.
- $\tilde{\psi}$ is asymptotically efficient and doubly robust.
- If each subject interferes with $\leq K$ other subjects, as in van der Laan (2014), then

$$\sqrt{n}(\tilde{\psi} - \psi) \rightarrow N(0, var(IF))$$

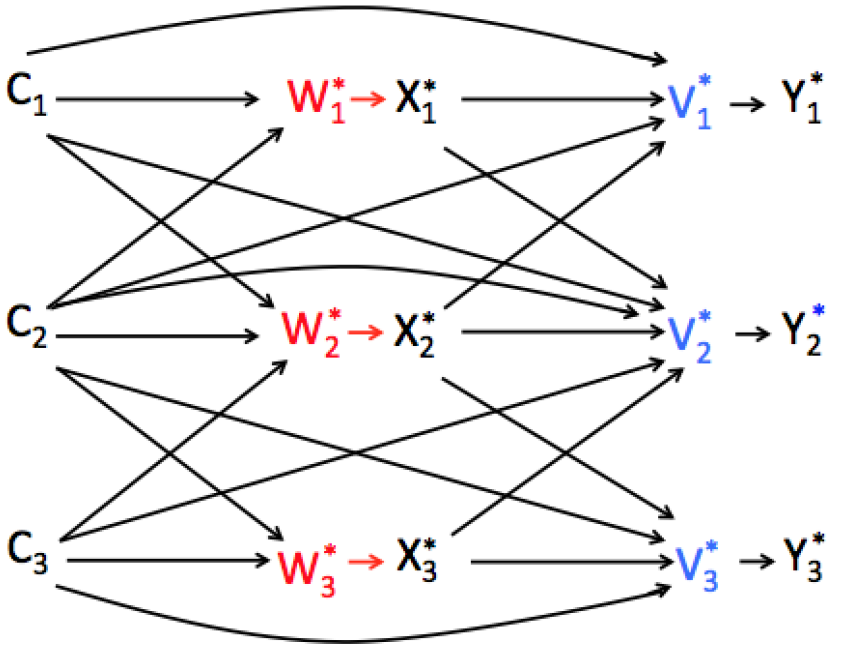
- Instead, we let $K_n \rightarrow \infty$ as $n \rightarrow \infty$ s.t. $\frac{K_n^2}{n} \rightarrow 0$. Then

$$\sqrt{C_n}(\hat{\psi} - \psi) \rightarrow N(0, var(IF)),$$

where $\frac{n}{K_n^2} \leq C_n \leq n$.

stochastic network interventions

- We can also identify the effects of **stochastic** interventions that replace f_X with a new, user-specified distribution:



- For each x in the support of X, X_i is set by the intervention to x with probability given by the stochastic intervention distribution.

stochastic network interventions

- Examples include
  - interventions that add, remove, or relocate ties in the network.
  - interventions that change the dependence of a subject's treatment on other subjects' covariates, or of a subject's outcome on other subjects' covariates and treatments.
  - Interventions on summary features of network topology:
    - An intervention on features of the network topology replaces $\mathbf{T}$ with the members of a class $\mathscr{T}^*$ of $n \times n$ adjacency matrices that share the intervention features, stochastically according to some probability distribution $g_{\mathbf{T}^*}$ over $\mathscr{T}^*$.
    - Whether or not we can define, identify, and estimate interventions involving these features of network topology hinges crucially on the positivity assumption.
    - e.g. degree / centrality

principled approach

- Pros
  - uses all of the available data
  - estimands are unconditional
  - efficient and doubly robust estimation
- Cons
  - hard(er) to understand, hard to implement
  - may not be clear in finite samples what to do with K and with hubs

For details see Ogburn et al, Causal inference for social network data (available on arXiv)

summary and next steps

- Although it is accepted practice in many areas, it can be very dangerous to assume that observations are independent when they may not be!
- When it's available, we can use the information barrier structure to facilitate inference even when subjects are connected in complicated ways.
- Future work is needed to adapt results from spatial statistics to deal with non-independence of observations.
  - This is necessary for latent variable dependence.
  - It is desirable when network dependence is due to contagion, because it permits inference from more realistic/feasible data structures.

Thank you

Why can't we use spatial dependence results?

- Network topology doesn't naturally correspond to Euclidean space.
  - In order to embed a network in $\mathbb{R}^d$, we would have to let d grow with sample size.
  - Spatial results require d to be fixed.
- Population growth is usually assumed to occur at the boundaries of the d-dimensional space.
  - It's not clear how to define boundaries in networks.
- Mixing assumptions and m-dependence don't imply bounded correlation structure.
  - In spatial data most observations are distant from one another.
  - The maximum network-based distance between two observations may be very small.
  - The distance distribution may not be right-skewed enough.