Vaccines, Contagion, and Social Networks

Elizabeth L. Ogburn*, Tyler J. VanderWeele**

*Department of Biostatistics, Johns Hopkins University,
**Program on Causal Inference, Harvard University

outline

- Brief history of causal inference about contagious processes in social networks
- Brief intro to infectious disease setting
- Interlude / punchline
- Definition of contagion & infectiousness effects
- Details
  - identification
  - estimation in independent groups
  - estimation in social networks

brief history 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 of causal inference using network data

- Statisticians have responded critically to this approach (e.g. Cohen-Cole and Fletcher, 2008; Lyons, 2011; Noel and Nyhan, 2011; Shalizi and Thomas, 2011):
  - Shalizi & Thomas (2011) explained that homophily and peer effects are often impossible to disentangle; conditioning on the alter's past doesn't suffice to control for homophily.
  - Lyons (2011) pointed out model incoherence related to overidentification due to including each observation in more than one regression equation and as both predictor and outcome
- VanderWeele et al (2012) demonstrated that these models could be correctly specified and coherent under the null hypothesis of no peer effects and no other sources of dependence across subjects.
  - But we proved that they give anticonservative standard errors under H_A.
  - We proposed using these models for hypothesis testing – but what is H_0?
  - Shalizi (2012) noted problems with power...

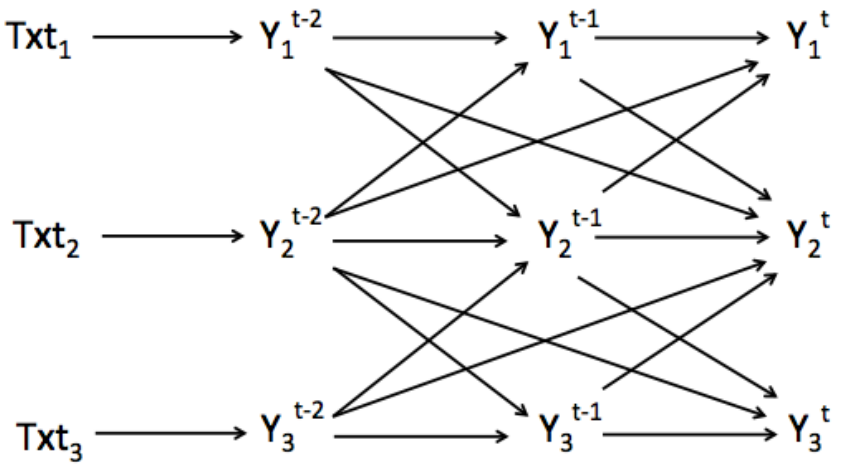
brief history causal inference using network data

- Randomization based inference facilitates hypothesis testing (e.g. Toulis & Kao, 2013; Bowers et al., 2013).
- Work on interference generally relies on randomization and/or multiple independent groups (e.g. Sobel, 2006; Hong & Raudenbush, 2006; Rosenbaum, 2007; Hudgens & Halloran, 2008; Tchetgen Tchetgen & VanderWeele, 2012; Aronow & Samii, 2013).
- 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).

- For many infectious diseases, homophily is probably not an issue.
  - Infection is not related to latent traits that could be the basis of friendship.
  - Timing of infection is easy to observe, making it easy to control for.
  - Contrast with, e.g., obesity...
- All of the dependence in the system comes from a contagious process and can therefore be observed and, in a sense, conditioned away.

- interlude -

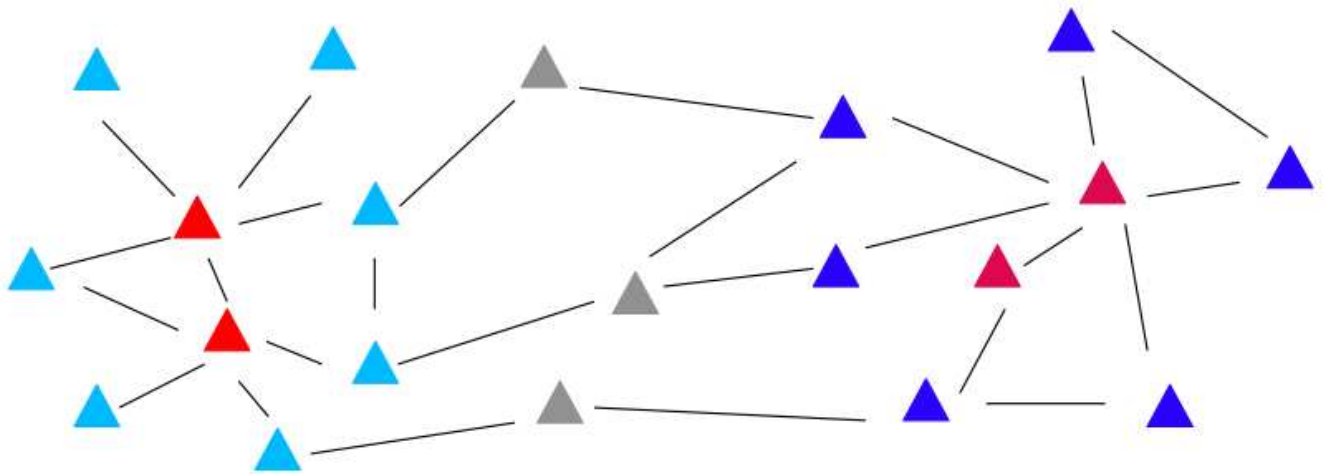
- When dependence is due **solely** to a contagious process, it implies information barrier structures, e.g.
$\left[Y_1^t \perp Y_2^t \mid Y_1^{t-2}, Y_2^{t-2}, Y_1^{t-1}, \text{ and } Y_2^{t-1}\right]$ and $\left[Y_1^{t-2} \perp Y_3^{t-1}\right]$.



- 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.

create conditionally independent units

- Randomly sample non-overlapping groups from the network.



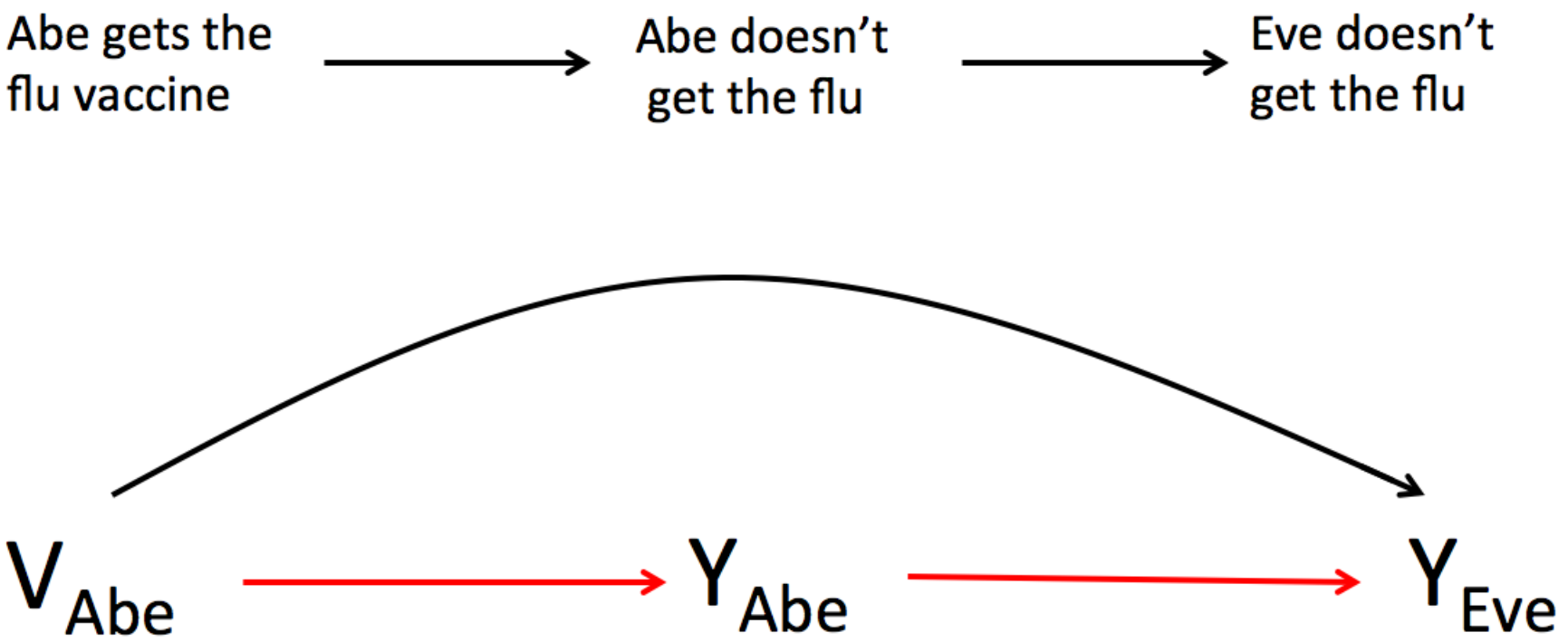
- This will allow us to condition on an "information barrier."
- Now can estimate conditional estimands using standard statistical machinery like GLMs.
  - The residuals will be uncorrelated across subjects despite the dependence structure.

- end interlude -

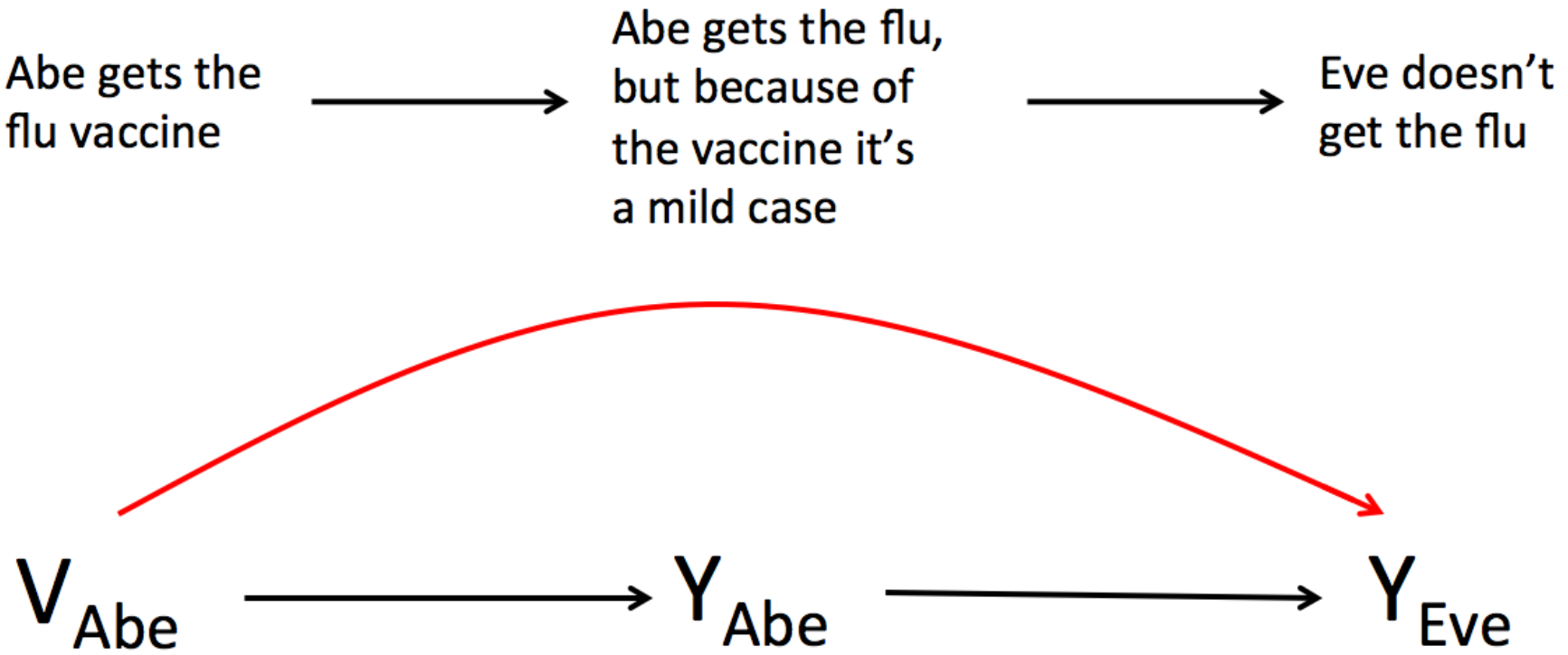
contagion v. infectiousness

- Suppose Abe and Eve are friends. If Abe is vaccinated against the flu at the beginning of flu season, this could have a protective effect on Eve – that is, the **total effect of V_{Abe} on Y_{Eve}** is negative (protective) and significant.
- But how does this effect operate?
- Two possible mechanisms: contagion and infectiousness (VanderWeele et al, 2012)

the contagion effect



the infectiousness effect



identification

- VanderWeele et al. (2012) considered identification and estimation of the infectiousness and contagion effects in simple setting:
- Sample comprised of independent households of size two with one member of each household assumed to be homebound.
- Next up: sample of independent households, but both individuals may be exposed outside the household (Ogburn & VanderWeele, 2013; Ogburn et al., 2014).

- VanderWeele et al. defined the contagion and infectiousness effects as the **natural indirect and natural direct effects** of Abe's vaccination on Eve's flu status, with Abe's flu status as mediator.
- But it's not always that simple! Considerations:
  - These effects are only operational if Abe would get the flu before Eve if he were not vaccinated.
  - The infectiousness effect can only operate if Abe gets the flu.
  - The flu has a time line – if Eve gets the flu weeks after Abe, she probably didn't catch it from him.

- Let T be the time of the first case of the flu between Abe and Eve.
- Define $T(v)$ to be the counterfactual time of the first flu if Abe had vaccine status $v \in (0,1)$.
- Let s be the sum of the incubation and infectiousness periods for the flu. If Abe transmits the flu to Eve, it has to happen by time $T+s$.
- Let $Y_e^{T(v')+s}\left(v, Y_a^{T(v')}(v')\right)$ be the counterfactual flu status we would have observed for Eve at time $T(v')+s$ if Abe's vaccine status were set to v and his flu status at time $T(v')$ were set to its counterfactual under vaccine status v'.

- The contagion effect is given by

$$\frac{Y_e^{T(1)+s}\left(0, Y_a^{T(1)}(1)\right)}{Y_e^{T(0)+s}\left(0, Y_a^{T(0)}(0)\right)}$$

- The infectiousness effect is given by

$$\frac{Y_e^{T(1)+s}\left(1, Y_a^{T(1)}(1)\right)}{Y_e^{T(1)+s}\left(0, Y_a^{T(1)}(1)\right)}$$

- The product of these two effects is the total effect of Abe's vaccination on Eve's flu status:

$$\frac{Y_e^{T(1)+s}(1)}{Y_e^{T(0)+s}(0)}$$

- We can never identify these effects for Abe and Eve specifically, but we can hope to identify the average contagion effect

$$Con = \frac{E\left[Y_e^{T(1)+s}\left(0, Y_a^{T(1)}(1)\right)\right]}{E\left[Y_e^{T(0)+s}\left(0, Y_a^{T(0)}(0)\right)\right]}$$

and the average infectiousness effect

$$Inf = \frac{E\left[Y_e^{T(1)+s}\left(1, Y_a^{T(1)}(1)\right)\right]}{E\left[Y_e^{T(1)+s}\left(0, Y_a^{T(1)}(1)\right)\right]}.$$

- The product of these two effects is the average total effect of an alter's vaccination on an ego's flu status:

$$\frac{E\left[Y_e^{T(1)+s}(1)\right]}{E\left[Y_e^{T(0)+s}(0)\right]}$$

- The contagion and infectiousness effects are the **natural indirect and natural direct effects** of Abe's vaccination on Eve's flu status, with Abe's flu status as mediator.

- Identification requirements for natural direct and indirect effects are well known:

  1. No unmeasured treatment-mediator confounding,
  2. No unmeasured mediator-outcome confounding,
  3. No unmeasured treatment-outcome confounding,
  4. No post-treatment mediator-outcome confounder ("recanting witness").

- Let C be a collection of confounders that satisfy 1 through 3.
- If Eve may be vaccinated, then V_e should be included in C.
- If Abe and Eve have mutual friends or contacts, those people are potential confounders of Abe and Eve's flu statuses (i.e. of the mediator-outcome relationship).
- As long as they are observed, we can include them in C.
  - e.g. $\sum_i Y_i^{T-b}$, where b is the incubation period.
  - The rational is that $T-b$ is the latest time at which any flu activity among mutual contacts could be a cause of the mediator and therefore a confounder of the mediator-outcome relationship.

estimation in independent groups

- Treating each group (Abe, Eve, friends & associates) as a single unit, this is no different from estimation of natural direct and indirect effects in other settings:
  - Requires models for $E\left[Y_e^{T+s} \mid V_a, Y_a^T, \boldsymbol{C}\right]$ and $E\left[Y_a^T \mid V_a, \boldsymbol{C}\right]$.
- One important exception: unlike in other settings, Y_e^{T+s} is, by definition, equal to 0 whenever Y_a^T is equal to 0;
  - Any model specified for for $E\left[Y_e^{T+s} \mid V_a, Y_a^T, \boldsymbol{C}\right]$ must be consistent with this restriction.

- For example, if

$$\begin{aligned} log\left\{E\left[Y_e^{T+s} \mid V_a, Y_a^T = 1, \boldsymbol{C}\right]\right\} &= \gamma_0 + \gamma_1 V_a + \gamma_2' \boldsymbol{C} \\ logit\left\{E\left[Y_a^T \mid V_a, \boldsymbol{C}\right]\right\} &= \eta_0 + \eta_1 V_a + \eta_2' \boldsymbol{C} \end{aligned}$$

then the contagion effect

$$\frac{E\left[Y_e^{T(1)+s}(0, Y_a^{T(1)}(1)) \mid \boldsymbol{c}\right]}{E\left[Y_e^{T(0)+s}(0, Y_a^{T(0)}(0)) \mid \boldsymbol{c}\right]} = \frac{e^{\eta_1} + e^{\eta_0+\eta_1+\eta_2'\boldsymbol{c}}}{1 + e^{\eta_0+\eta_1+\eta_2'\boldsymbol{c}}}$$

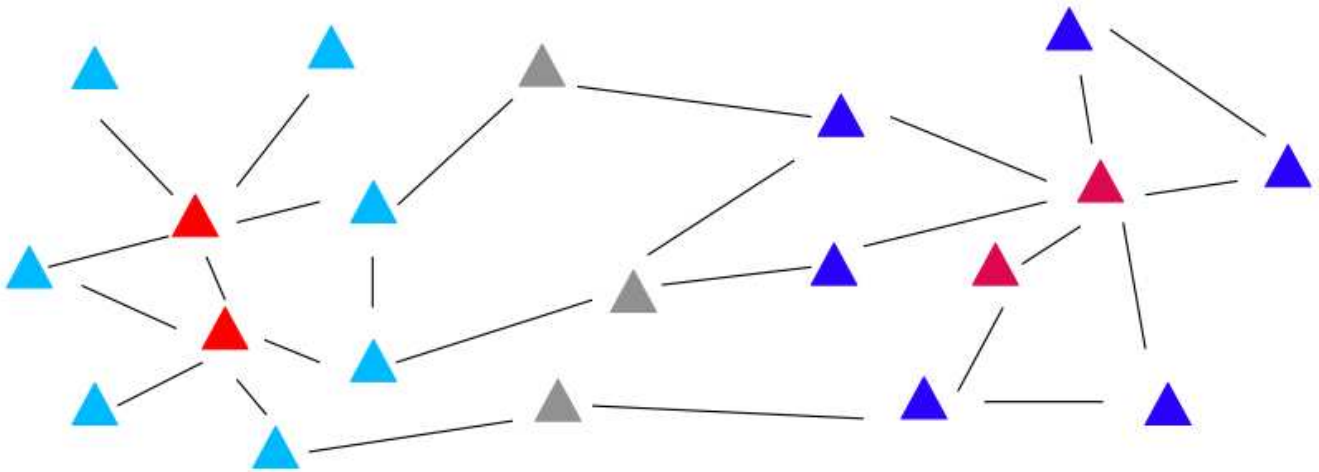
and the infectiousness effect

$$\frac{E\left[Y_e^{T(1)+s}(1, Y_a^{T(1)}(1)) \mid \boldsymbol{c}\right]}{E\left[Y_e^{T(1)+s}(0, Y_a^{T(1)}(1)) \mid \boldsymbol{c}\right]} = e^{\gamma_1}.$$

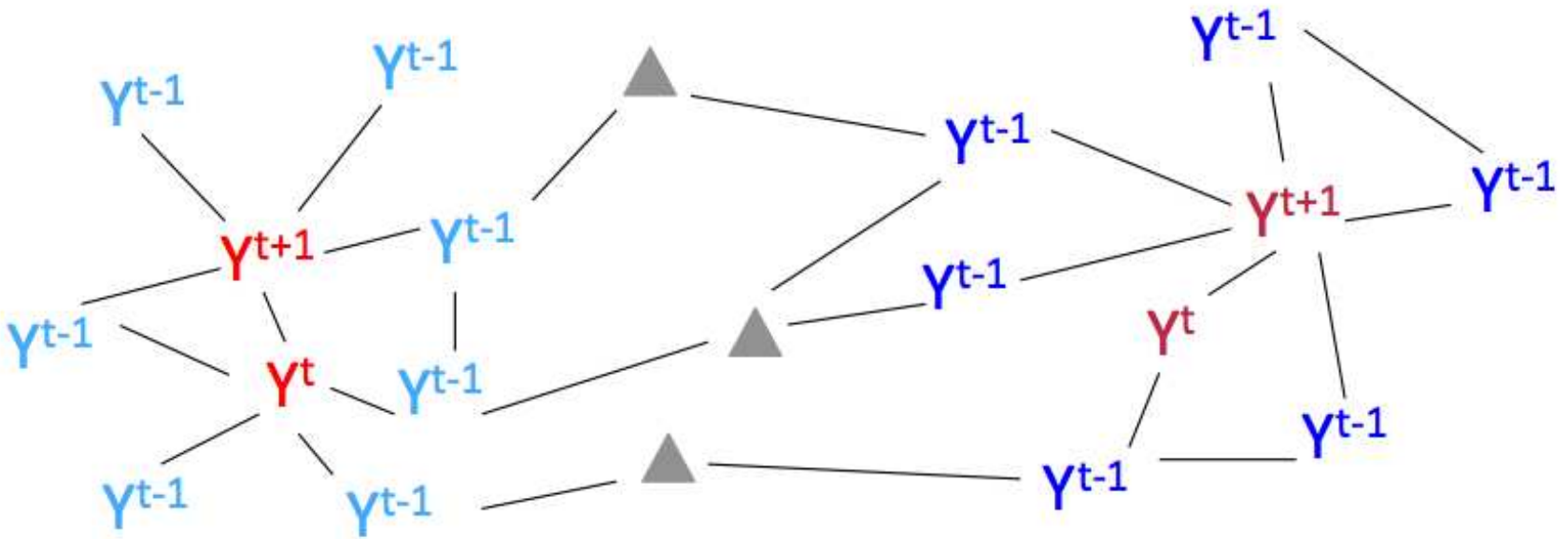
contagion and infectiousness effects in a network

- If an infectious disease is observed and treated in a single interconnected community or social network, then inference is much more difficult.
- It is desirable for a number of reasons to study infectiousness and contagion in the context of social networks rather than in independent communities.
  - e.g. less costly to collect the data, corresponds better to the true nature of vaccine programs...
- Extending to the network setting is not a challenge for identification, but it is a challenge for estimation.

- Randomly sample non-overlapping groups from the network.



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


Y t-1
Y t-1
Y t-1
Y t-1
Y t-1
Y t+1
Y t-1
Y t-1
Y t-1
Y t+1
Y t-1
Y t
Y t
Y t-1
Y t-1
Y t-1
Y t-1

details

- ***Step 1***
  Select K pairs of nodes such that the two nodes in each pair share a tie, but, for each pair, neither node nor any of their contacts has a tie to a node in any other pair or to the contacts of any member of any other pair. Randomly select one member of each pair to be the ego and one to be the alter.

- ***Step 2***
  Let $U_{e_k}^{T_k+f}$ and $L_{e_k}^{T_k+f}$ be the number of unvaccinated and vaccinated nodes, respectively, with ties to e_k who were sick by time $T_k + f$. Define $U_{a_k}^{T-b}$ and $L_{a_k}^{T-b}$ similarly as the number of unvaccinated and vaccinated nodes, respectively, with ties to a_k who were sick by time $T_k - b$.

- ***Step 3***
  Estimate

$$Con^* = \frac{E\left[Y_e^{T(1)+s}(0, Y_a^{T(1)}(1)) \mid U_a^{T-b}, L_a^{T-b}, U_e^{T+f}, L_e^{T+f}, \boldsymbol{C}\right]}{E\left[Y_e^{T(0)+b}(0, Y_a^{T(0)}(0)) \mid U_a^{T-b}, L_a^{T-b}, U_e^{T+f}, L_e^{T+f}, \boldsymbol{C}\right]}$$

$$Inf^* = \frac{E\left[Y_e^{T(1)+s}(1, Y_a^{T(1)}(1)) \mid U_a^{T-b}, L_a^{T-b}, U_e^{T+f}, L_e^{T+f}, \boldsymbol{C}\right]}{E\left[Y_e^{T(1)+s}(0, Y_a^{T(1)}(1)) \mid U_a^{T-b}, L_a^{T-b}, U_e^{T+f}, L_e^{T+f}, \boldsymbol{C}\right]}$$

  (and standard errors).

- Fit two models:

$$g(E\left[Y_{e_k}^{T_k+s} \mid V_{a_k}, Y_{a_k}^{T_k}=1, U_a^{T-b}, L_a^{T-b}, U_e^{T+f}, L_e^{T+f}, \boldsymbol{C}_k\right])$$
$$=\beta_0+\beta_1 V_{a_k}+\beta_2 U_a^{T-b}+\beta_3 L_a^{T-b}+\beta_4 U_e^{T_k+f}+\beta_5 L_e^{T_k+f}+\beta_6' \boldsymbol{C}_k$$

and

$$m(E\left[Y_{a_k}^{T_k} \mid V_{a_k}, U_a^{T-b}, L_a^{T-b}, U_e^{T+f}, L_e^{T+f}, \boldsymbol{C}_k\right])$$
$$=\alpha_0+\alpha_1 V_{a_k}+\alpha_2 U_a^{T_k-b}+\alpha_3 L_a^{T_k-b}+\alpha_4 U_e^{T_k+f}+\alpha_5 L_e^{T_k+f}+\alpha_6' \boldsymbol{C}_k.$$

- Under some assumptions, the residuals for these models will be uncorrelated across units, and therefore the resulting GLMs will be correctly specified and give valid standard errors, under the null and the alternative.

conclusion and next steps

- This procedure has low power, but it is powered for the alternative hypothesis of interest.
- van der Laan (2012) and Ogburn and van der Laan (in progress) use the information barrier structure more efficiently and avoid throwing away information from some subjects.
- Adapt results from spatial statistics to deal with non-independence of observations.
  - This is necessary when there may be unstructured network dependence.
  - It is desirable when network dependence is due solely to a contagious process, because it permits inference from less rich data.

Thank you

references

Rubin, D. B. (1990). On the application of probability theory to agricultural experiments . essay on principles. section 9. comment: Neyman (1923) and causal inference in experiments and observational studies. Statistical Science 5, 472–480.

Halloran, M. E. & Struchiner, C. J. (1995). Causal inference in infectious diseases. Epidemiology , 142–151.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? Journal of the American Statistical Association 101, 1398–1407.

Hong, G. & Raudenbush, S. W. (2008). Causal inference for time-varying instructional treatments. Journal of Educational and Behavioral Statistics 33, 333–362.

Vansteelandt, S. (2007). On confounding, prediction and efficiency in the analysis of longitudinal and cross-sectional clustered data. Scandinavian Journal of Statistics 34, 478–498.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. Journal of the American Statistical Association 102, 191–200.

Hudgens, M. G. & Halloran, M. E. (2008). Toward causal inference with interference. Journal of the American Statistical Association 103, 832–842.

references

Graham, B. S., Imbens, G. W. & Ridder, G. (2010). Measuring the effects of segregation in the presence of social spillovers: A nonparametric approach. Tech. rep., National Bureau of Economic Research.

Manski, C. F. (In press). Identification of treatment response with social interactions. The Econometrics Journal.

Tchetgen Tchetgen, E. J. & VanderWeele, T. J. (2012). On causal inference in the presence of interference. Statistical Methods in Medical Research 21, 55–75.

Aronow, P. & Samii, C. (2012). Estimating Causal EffectsUnder General Interference. Working paper (http://pantheon.yale.edu/~pma5/ate-interference.pdf).

van der Laan, M. J. (2012). Causal Inference for Networks. U.C. Berkeley Division of Biostatistics Working Paper Series Paper 300.

Lyons, R. (2011). The spread of evidence-poor medicine via flawed social network analyses. Statistics, Politics and Policy 2(1), Article 2, 1-26.

VanderWeele, T.J., Ogburn, E.L. & Tchetgen Tchetgen, E.J. (2012). Why and When "Flawed" Social Network Analyses Still Yield Valid Tests of no Contagion. Statistics, Politics, and Policy 3(1).

references

Christakis, N.A. and Fowler, J.H. (2011). Social contagion theory: examining dynamic social networks and human behavior. Statistics in Medicine, to appear.

Shalizi, C.R., Thomas, A.C. (2011). Homophily and contagion are generically confounded in observational social network studies. Sociological Methods and Research, 40:211-239.

Shalizi C.R. (2012). Comment on Why and When 'Flawed' Social Network Analyses Still Yield Valid Tests of no Contagion. Statistics, Politics, and Policy 3(1).

Ogburn, E.L. & VanderWeele, T.J. (2012). Causal diagrams for interference and contagion. Under revision.

Chen, LHY (1978). Two central limit problems for dependent random variables. Probability Theory and Related Fields 43, no. 3: 223-243.

Barbour, A. D., Karoński, M., & Ruciński, A. (1989). A central limit theorem for decomposable random variables with applications to random graphs. Journal of Combinatorial Theory, Series B, 47(2), 125-145.

Rinott, Y., & Rotar, V. (1996). A Multivariate CLT for Local Dependence with $n^{1/2} log\, n$ Rate and Applications to Multivariate Graph Related Statistics. Journal of multivariate analysis, 56, 333-350.

references

Raic, M. (2003). Normal approximation by Stein's method. In Proceedings of the seventh young statisticians meeting.

Chen, L. H., & Shao, Q. M. (2005). Stein's method for normal approximation. An introduction to Stein's method, 4, 1-59.

Chen, L. H., & Shao, Q. M. (2004). Normal approximation under local dependence. The Annals of Probability, 32(3), 1985-2028.

Ibragimov, R., & Müller, U. K. (2010). t-Statistic based correlation and heterogeneity robust inference. Journal of Business & Economic Statistics, 28(4), 453-468.

Bester, C. A., Conley, T. G., & Hansen, C. B. (2011). Inference with dependent data using cluster covariance estimators. Journal of Econometrics, 165(2), 137-151.

Lahiri, S. N. (2003). Resampling methods for dependent data. Springer.

Politis, Romano & Wolf, 1999

assumptions

-

$$Y_i^t \perp Y_j^r \mid \left\{ \sum_{m \in \mathscr{A}_i : V_m = v} Y_m^{t-b}, v = 0, 1 \right\}, \text{for all } j \notin \mathscr{A}_i \text{ and } r \leq t. \quad (1)$$

  - Contacts act as a causal barrier between two nodes who do not themselves share a tie. If two individuals, i and j, do not share a tie, then they can have no effect on one another's disease status that is not through their contacts' disease statuses.

- $$Y_i^t \perp V_j \mid \left\{ \sum_{m \in \mathscr{A}_i : V_m = v} Y_m^{t-b}, v = 0, 1 \right\} \text{ for all } j \notin \mathscr{A}_i \qquad (2)$$

and

$$Y_i^t \perp C_j \mid \left\{ \sum_{m \in \mathscr{A}_i : V_m = v} Y_m^{t-b}, v = 0, 1 \right\} \text{ for all } j \notin \mathscr{A}_i. \qquad (3)$$

  - Any effect of the covariates (including vaccination) of nodes without ties to i on i's disease status would again have to be mediated by the disease statuses of i's contacts. The infectiousness effect is not transitive: whether individual j caught the flu from a vaccinated or unvaccinated person has no influence on whether individual j transmits the flu.

- Embedded in these is the assumption that all ties are equivalent and all non-ties are equivalent with respect to transmission of the outcome.
- If there exists a person with whom two individuals in the network interact regularly, then that person is also in the network (with ties to both individuals). In some settings it may be possible to satisfy this condition, e.g. in full sociometric studies conducted de novo, or in studies of online data.