Causal Inference for Interference and Social Networks: Challenges and Tools

Elizabeth L. Ogburn

Department of Biostatistics,
Johns Hopkins University

introduction

- I will talk about causal inference when subjects interact with one another and outcomes are interdependent.
- Two distinct but related settings:
  - Interference – one subject's exposure can affect other subjects' outcomes.
  - Social networks – close social contacts have correlated outcomes.
- Two distinct challenges:
  - Nonparametric identification of causal effects.
  - Valid statistical inference taking into account a new kind of dependence.

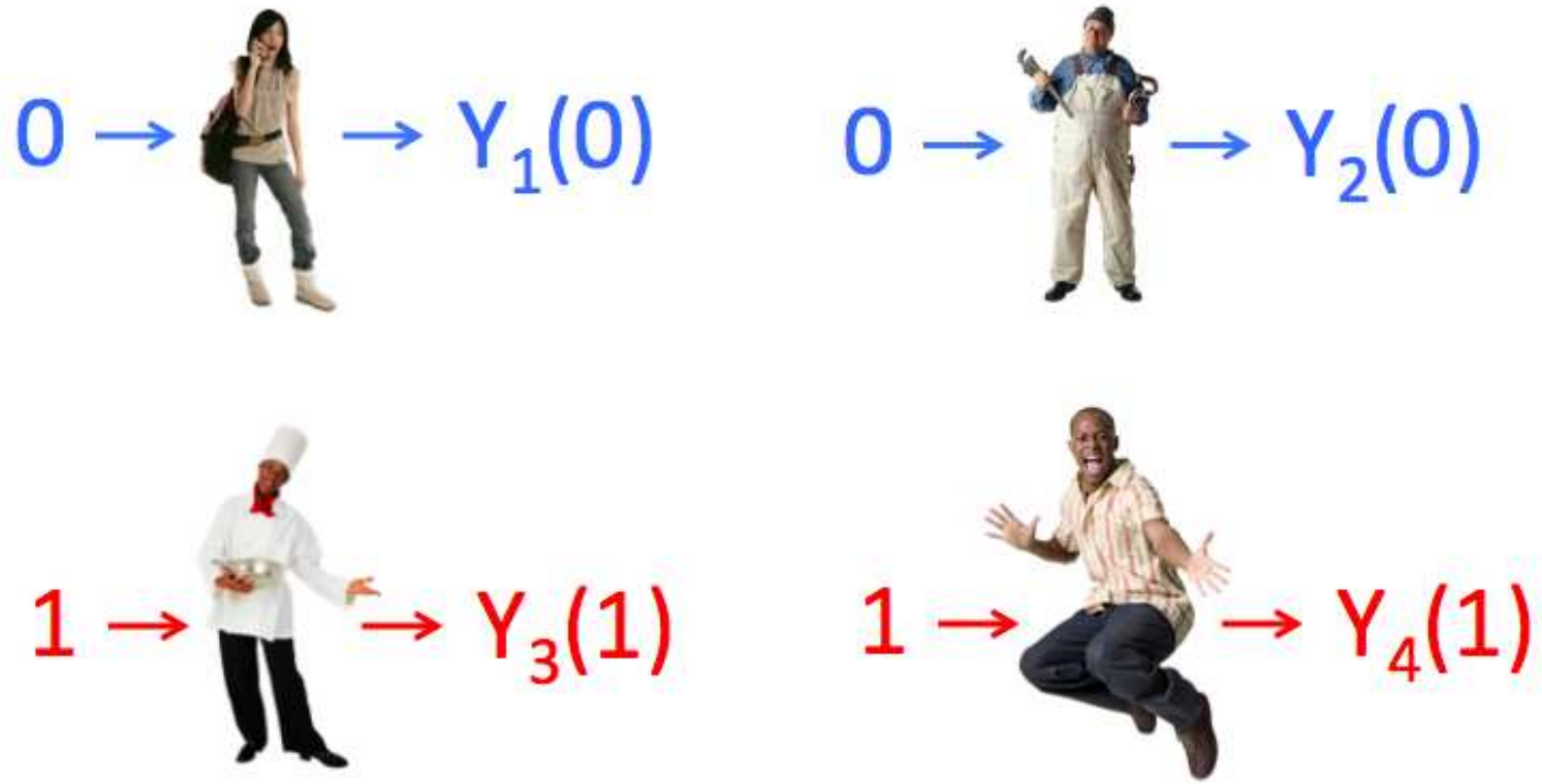
outline

- Background
- HopeNet study
- Interference
  - three distinct types of interference
  - nonparametric identification of causal effects using DAGs
- Social networks
  - statistical inference

iid data

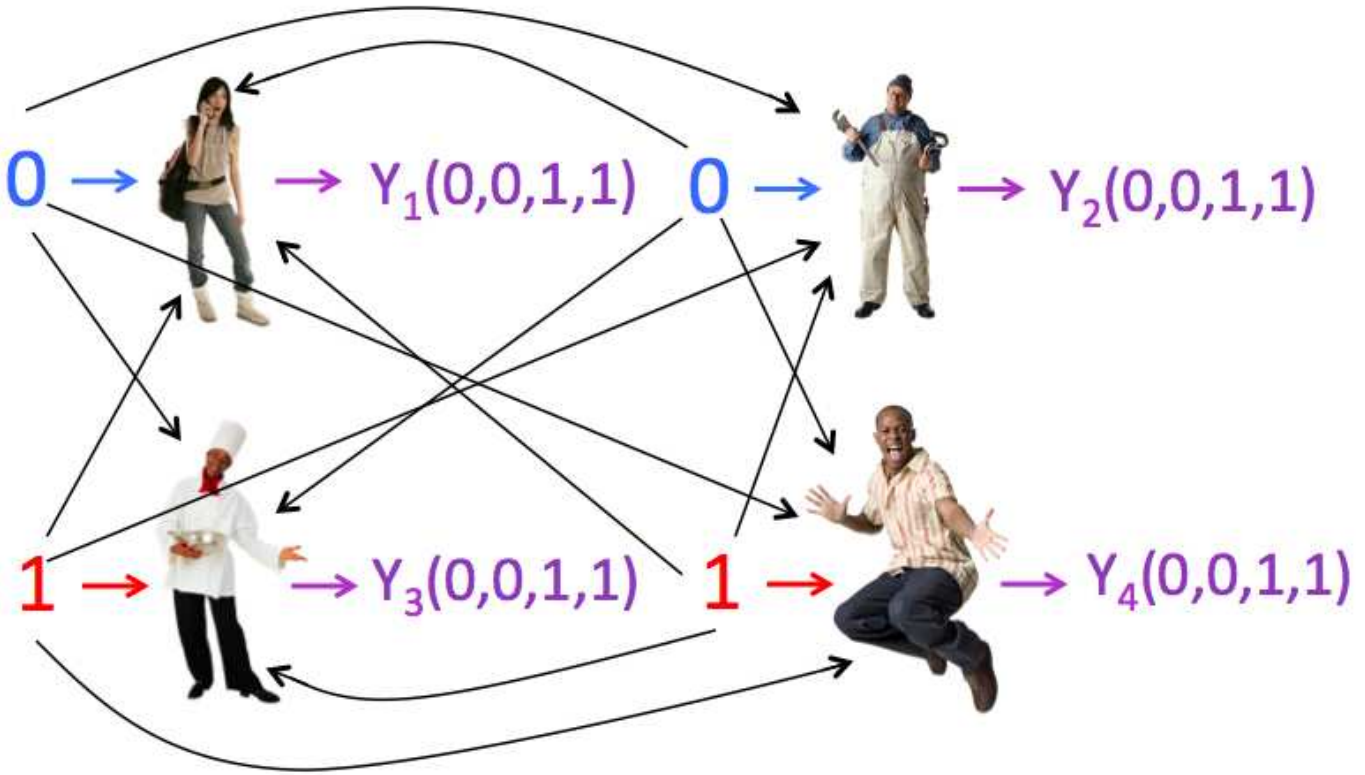
$$E[Y(1)] - E[Y(0)]$$

iid data



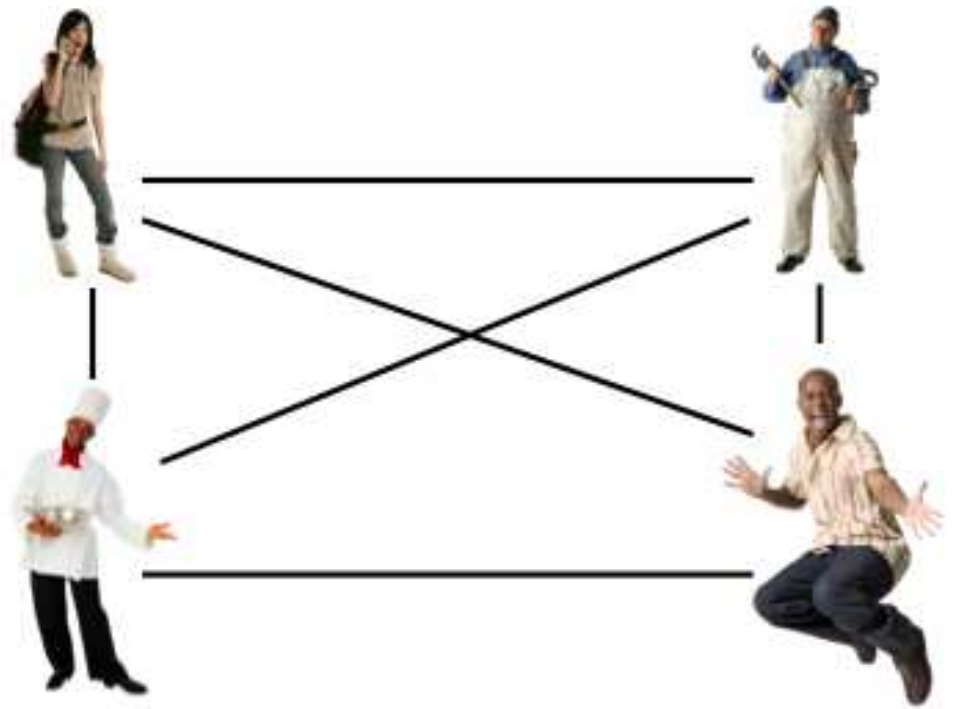
$$E[Y(1)] - E[Y(0)]$$

interference



- Under interference counterfactuals like $Y_i(a_i)$ are not well-defined.
- Instead, causal effects are contrasts of counterfactuals like $Y_i(\boldsymbol{a})$, where $\boldsymbol{a} = (a_1, \ldots, a_n)$.

social networks



- Contagion / influence / peer effects
- Correlated outcomes
- Thorny topology

HopeNet Study

- Health outcomes, progressive entrepreneurship, and Networks
- Nyakabare Parish, Mbarara Province, SW Uganda
  - 8 villages
  - 2000 adults

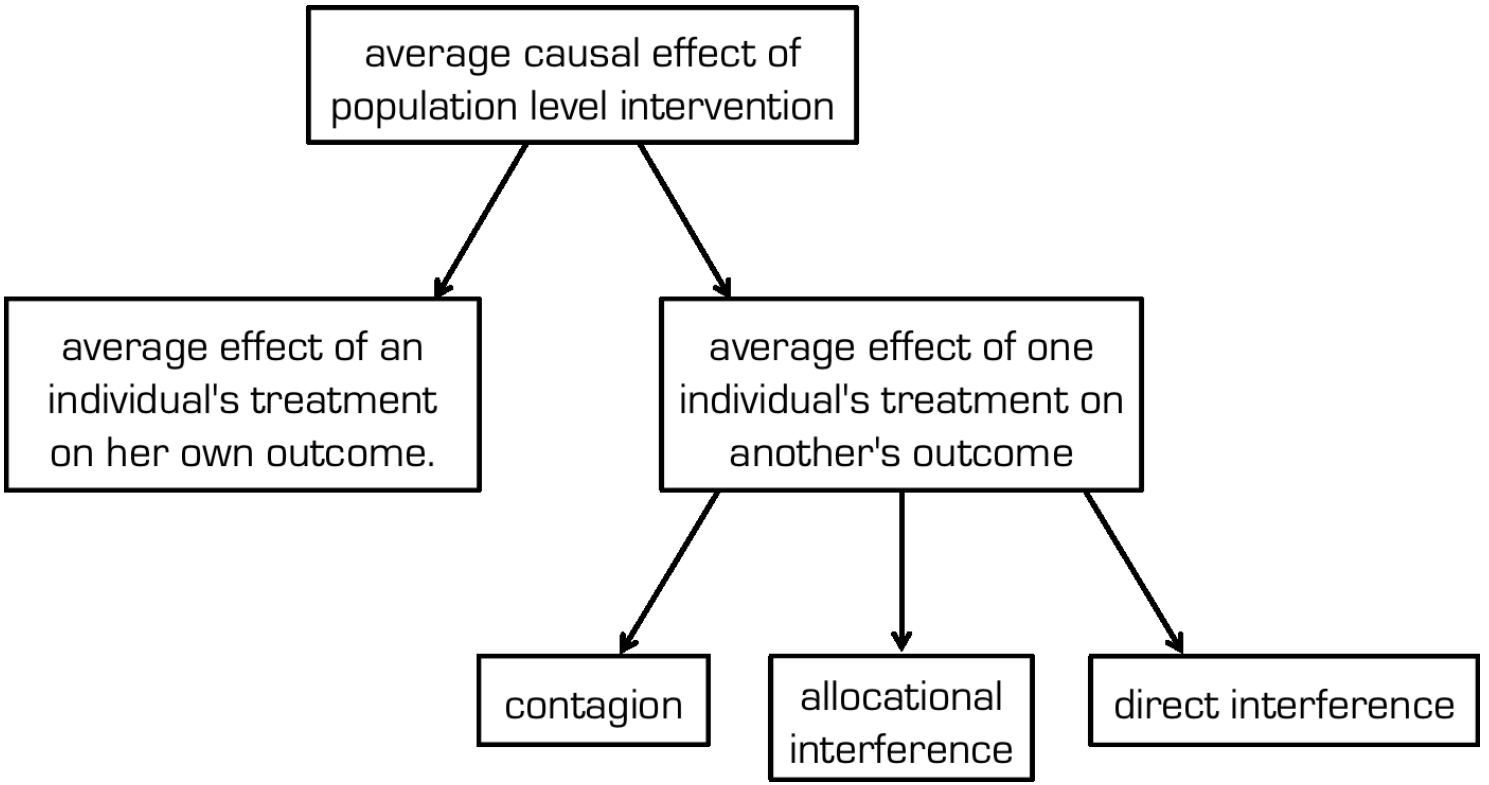
HopeNet Study

- Complete social network census for the adult population of the parish.
- Clean water intervention.

- Microenterprise intervention.

effects of interest

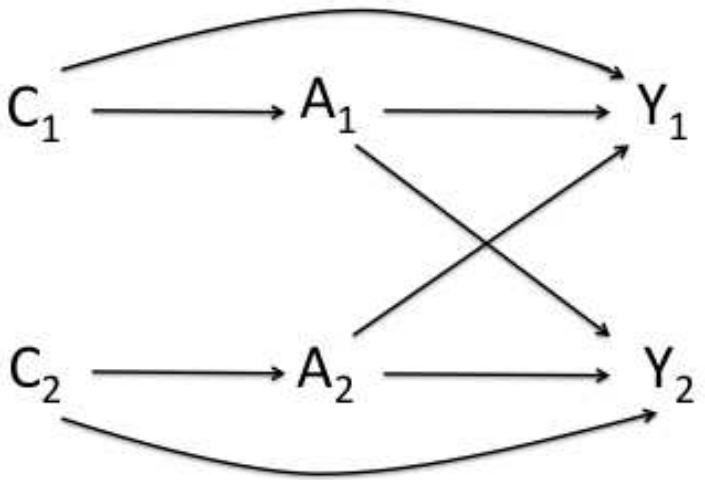
What types of effects are we interested in?



- Under what conditions are these effects nonparametrically identifiable?
  - Ogburn & VanderWeele (2012)
- Under what conditions can we perform inference about them?

direct interference

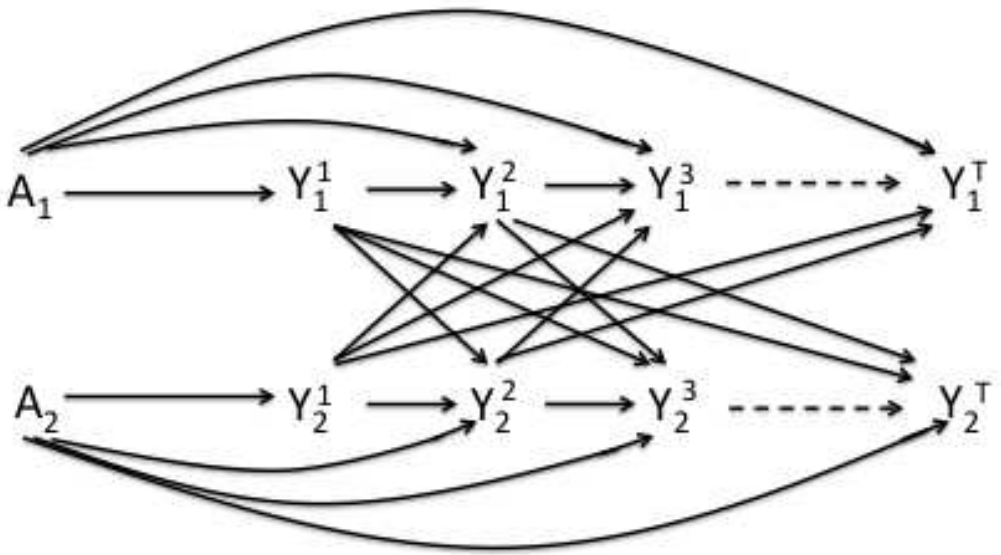
There is a causal pathway from one subject's treatment to another subject's outcome, not mediated by the first subject's outcome.



- This graphical representation suffices for all kinds of interference when we don't care *how* one subject's treatment affects another's outcome.
- Interference can be direct with respect to a particular outcome but not another, depending on how the outcomes are defined.

interference by contagion

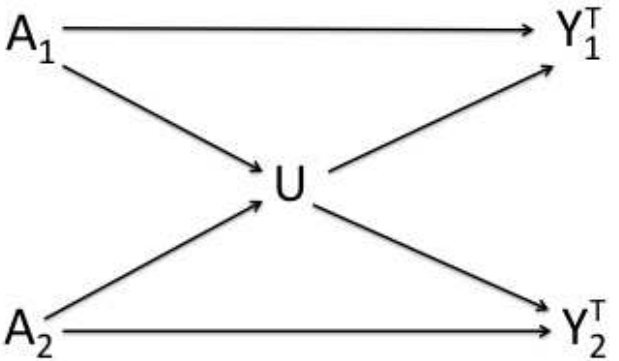
- The effect of one subject's treatment on another's outcome is mediated by the first subject's outcome.



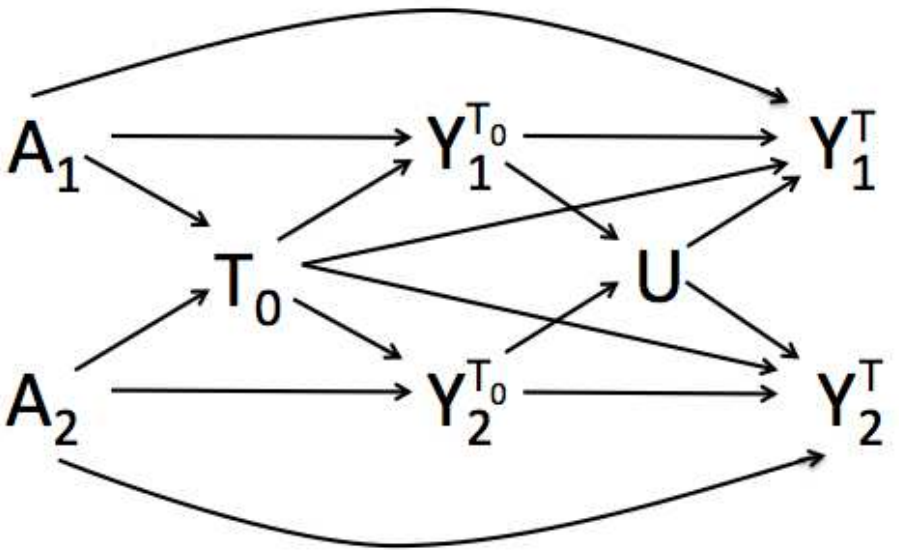
- Interference by contagion involves feedback among subjects' outcomes over time. (Two outcomes measured at the same time cannot causally affect one another.)
- The dotted arrows represent the intervening time points.

interference by contagion

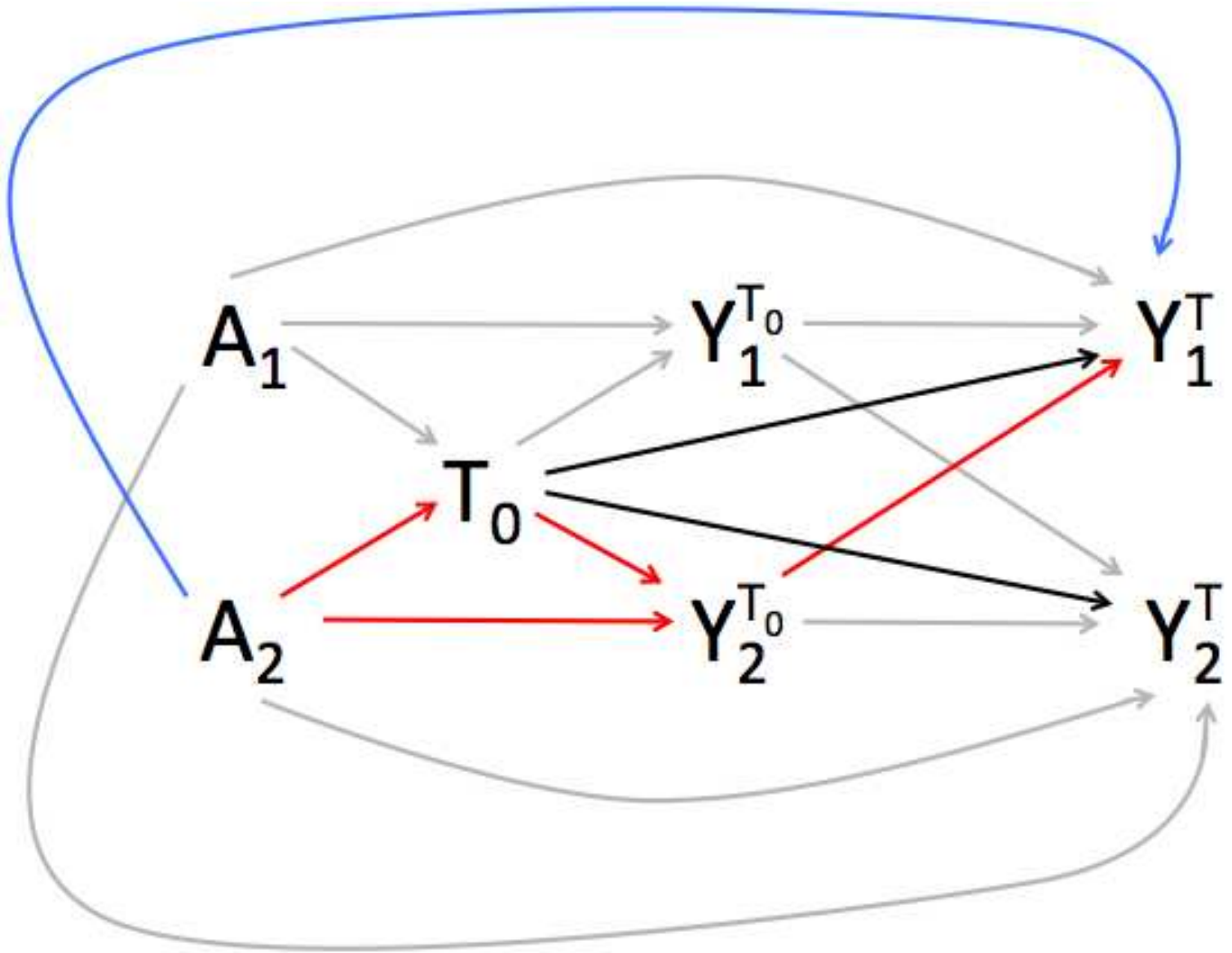
- If the outcome is observed only at time T (the end of follow up), then the operative DAG is



- If, in addition, the time T_0 of the first incident is observed for a binary outcome, then this is the operative DAG:

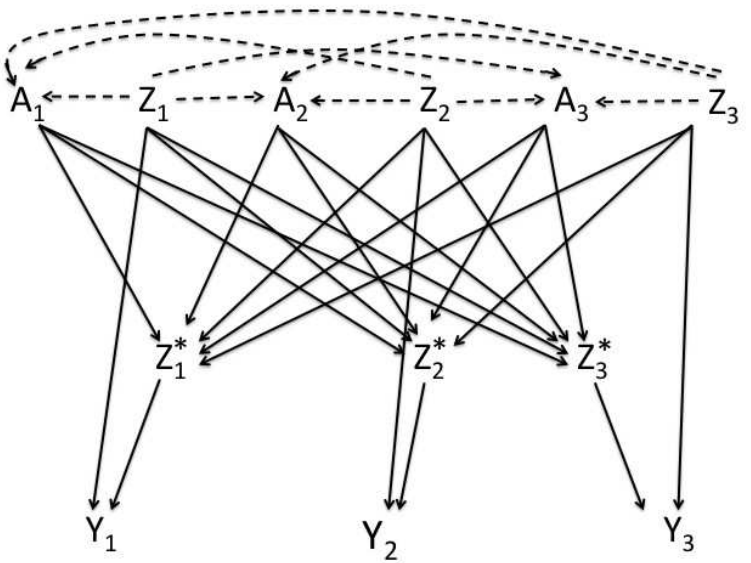


contagion or direct interference?



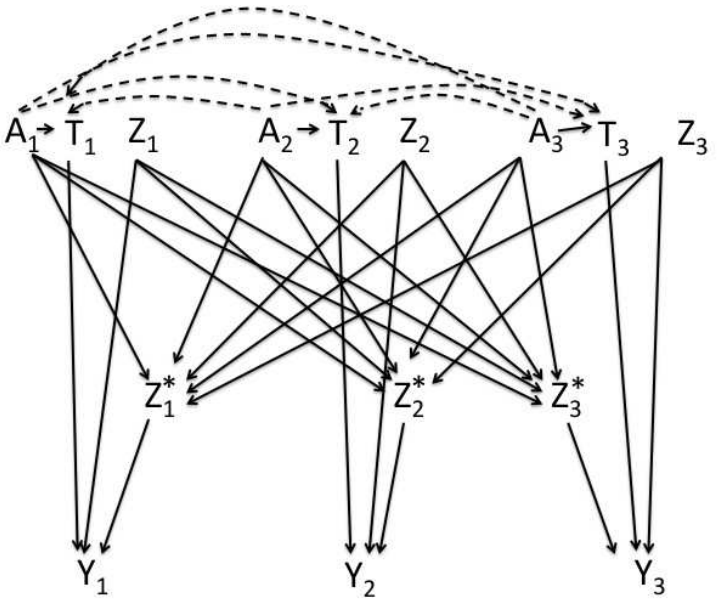
allocational interference

- Treatment in this setting allocates subjects to groups. Through their group-based interactions, subjects' characteristics may affect one another.



- A_i is a categorical variable indicating group assignment.
- Z_i is a vector of baseline covariates.
- $Z_i^* = Z \times I\{A = A_i\}$ is an array of the baseline covariates of all subjects assigned to the same group as subject i.

allocational interference



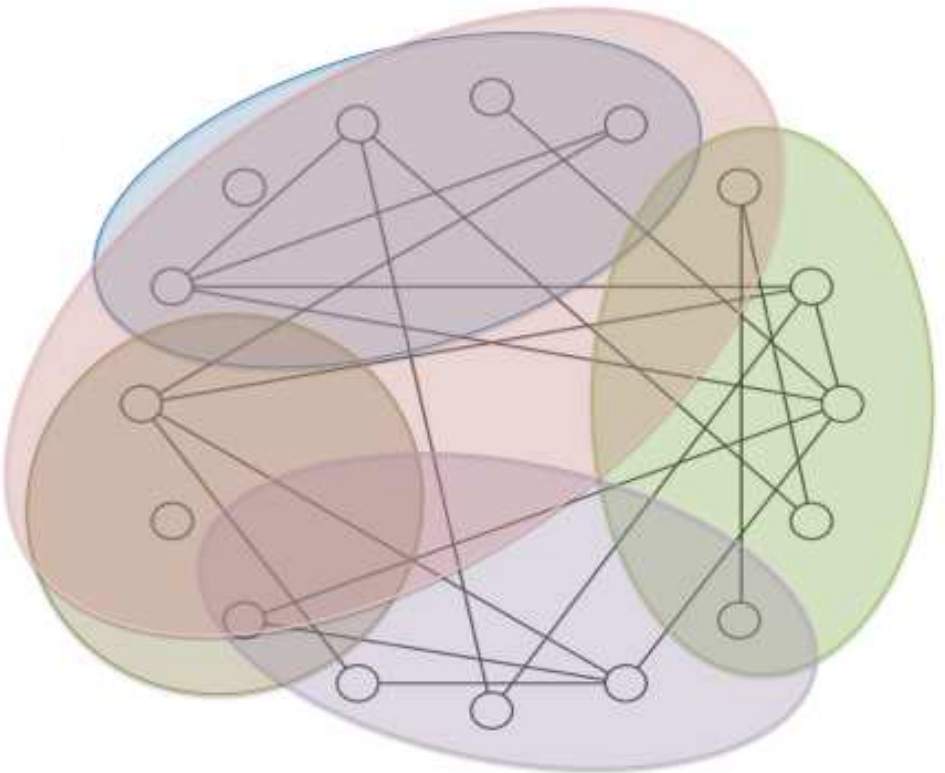
- If the groups are distinguished by properties other than their composition, we add to the DAG a new variable T_i , which is the group-level property to which subject i is exposed.
- If these properties are affected by group composition, we require arrows from Z into T.

inference

- Lots of work on inference for interference when independent blocks are observed. (Cf. Rubin, 1990; Halloran & Struchiner, 1995; Sobel, 2006; Hong & Raudenbush, 2006; Vansteelandt, 2007; Rosenbaum, 2007; Hudgens & Halloran, 2008; Graham et al., 2010; Manski, 2010; Tchetgen Tchetgen & VanderWeele, 2012)
- What about an interconnected network?
  - Nonparametric identification is principally the same;
  - Statistical inference can be quite different.

sources of network dependence

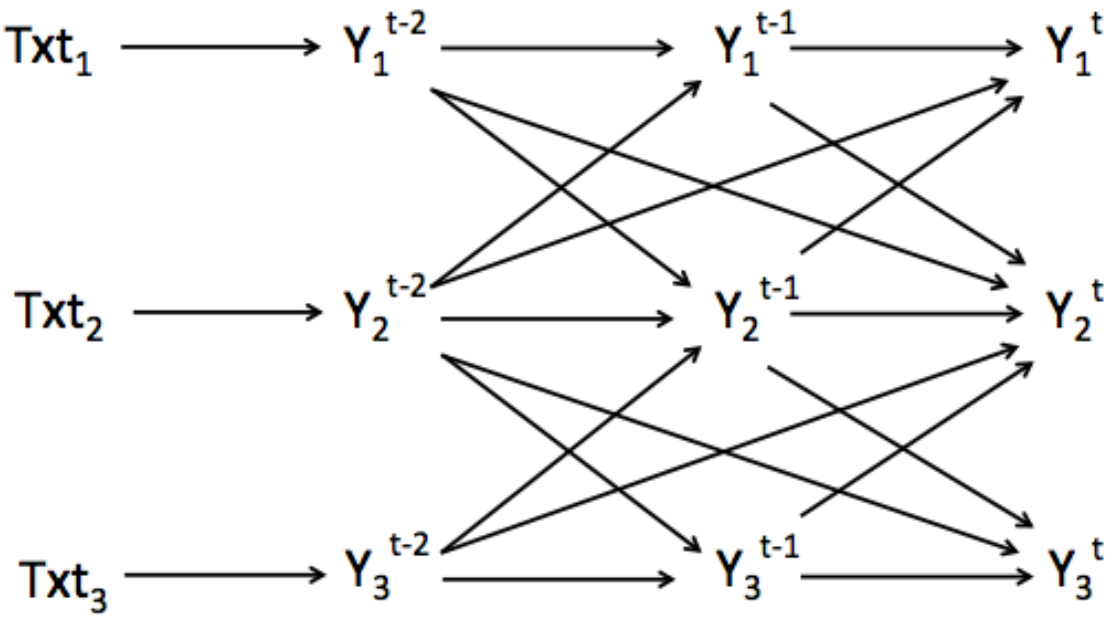
Latent variables cause outcomes among close social contacts to be more correlated than among distant contacts. (E.g. homophily, geography, shared culture.)



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.

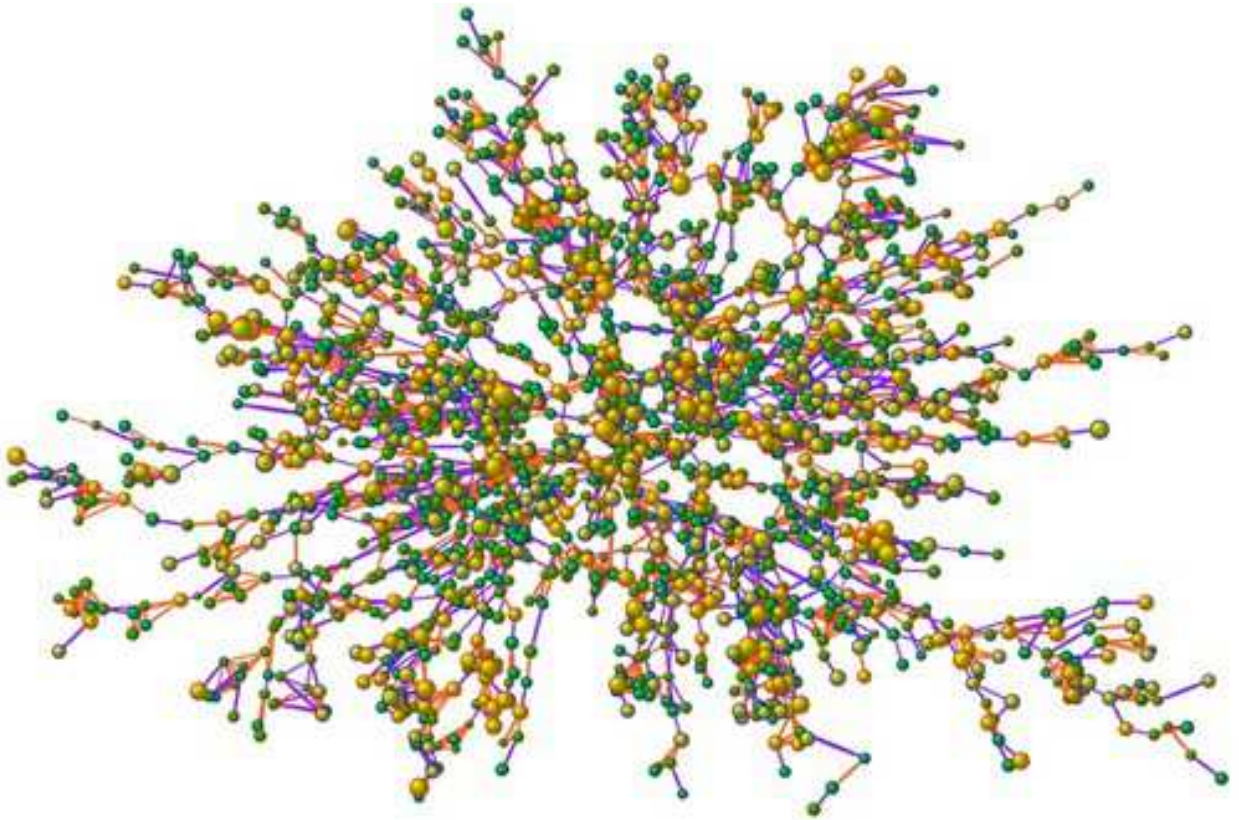
inference with network dependence

- Limitations of methods for spatial dependence.
- Some methods for statistical inference in networks.
  - local dependence
  - weakly dependent clusters
  - subsampling
  - k-dependence

definitions

- Stationarity: features of the distribution of observations does not depend on location in the network.
- Distance $\|i,j\|$ between nodes i and j: length of shortest path connecting i and j.
  - Other definitions are possible, e.g. taking into account number of paths or average path length between i and j.
- M-dependence: $W_i \perp W_j$ if $\|i,j\| > m$.
- Mixing conditions: $Cov(W_i, W_j) \to 0$ as $\|i,j\| \to \infty$.

Why can't we use spatial dependence results?



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 n.
  - Spatial results require d to be fixed or to grow slowly with n.
- 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 (nor how to define population growth).
- 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.

inference with network dependence

- Focusing for now on a traditional frequentist inferential framework, the challenge of network dependence is consistent s.e. estimation.

- Our target parameter is a population mean, μ.
- The sample mean, $\bar{W} = \sum_{i=1}^{n} W_i$, is unbiased for μ.
- The problem is consistent estimation of the asymptotic variance $Avar(\bar{W})$.
- Agnostic about sources of dependence (contagion vs. latent variables).
- Population growth must preserve key features of the network and data.

inference with network dependence

- Focusing for now on a traditional frequentist inferential framework, the challenge of network dependence is consistent s.e. estimation.

- Our target parameter is a population mean, μ.
- The sample mean, $\bar{W} = \sum_{i=1}^{n} W_i$, is unbiased for μ.
- The problem is consistent estimation of the asymptotic variance $Avar(\bar{W})$.
- Agnostic about sources of dependence (contagion vs. latent variables).
- Population growth must preserve key features of the network and data.

local dependence

- Local dependence: for each observation W_i, there is a set of indices I such that $W_i \perp W_j$ for $j \in I^c$.
  - M-dependence is a special case.
- Using Stein's method, it is easy to show that CLTs hold for locally dependent data, under restrictions on the size of I. (Cf. Chen, 1978; Barbour, Karonski & Rucinski, 1989; Rinott & Rotar, 1996; Raic, 2002; Chen & Shao, 2005)
- What types of networks are consistent with the restrictions on I?

weakly dependent clusters

If K clusters are asymptotically mean independent from one another, there are two approaches we might consider:

1. $T-$distribution based confidence intervals (Ibragimov & Muller, 2010; Bester, Conley & Hansen, 2011).
    - Requires asymptotic normality and mean stationarity at the cluster level.
2. Bootstrap the weakly dependent communities.
    - Stationarity is required only at the cluster level.

In the spatial dependence literature mean independence is justified with conditions on the relative size of the boundaries and interiors of the clusters; growth in d dimensions uniformly.

- These conditions don't translate into the network setting...

subsampling

- Subsampling has been used in many spatial dependence contexts (cf. Lahiri, 2003; Politis, Romano & Wolf, 1999), but neither the implementation nor the conditions under which it is appropriate are immediately applicable to networks.
- Under mild stationarity and dependence conditions, we can subsample to estimate $Avar(\bar{W})$:
  1. Select B subsamples of "consecutive" observations.
  2. In each subsample, calculate the subsample variance estimator $\widehat{\sigma}_b^2$.
  3. Estimate $Avar(\bar{W})$ with the average of the subsample estimators: $\widehat{\sigma}_{\bar{W}}^2 = \frac{1}{B}\sum_{b=1}^{B}\widehat{\sigma}_b^2$.

k-dependence

In some settings it may be expedient to estimate $Cov\left(\underset{\sim}{W}\right)$ directly.

- K-dependence: $Cov(W_i, W_j) = \sigma_k$, where $k = \|i, j\|$.
- Under k-dependence, m-dependence, and mean stationarity, we can get an unbiased and consistent estimate of $Cov\left(\underset{\sim}{W}\right)$ by this procedure:
  1. For each $k < m$, select pairs of nodes that are k units apart, such that the pairs themselves are at least m units apart from on another.
  2. Estimate $\widehat{\sigma}_k$ with the average covariance across the selected pairs.
  3. Estimate $Cov\left(\underset{\sim}{W}\right)$ with the plug-in estimator.
- This doesn't demand as much from m-dependence as other procedures do...

other directions

- Different types of asymptotics:
  - combine infill and increasing domain asymptotics,
  - *fractal* or *tessellation* asymptotics.
- Identify low-level conditions for CLTs and LLNs in network settings.
- Learn a new, latent distance metric. (E.g. work by Adrian Raftery & others)
- Finite sample results using bounded influence:
  $Var(W_i) >> \sum_j Cov(W_i, W_j)$.

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 instruc- tional 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 inter- ference. 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 áawed 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.

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, L.H.Y. (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, D., Romano, J. P., & Wolf, M. (1999). Weak convergence of dependent empirical measures with application to subsampling in function spaces. Journal of statistical planning and inference, 79(2), 179-190.