Causal Estimation of Peer Effects Using Instrumental Variables

James O'Malley, Ph.D.

The Dartmouth Institute, Geisel School of Medicine, Dartmouth College
email: James.OMalley@Dartmouth.edu

SAMSI, August 21, 2013

Acknowledgements: Felix Elwert, Niels Rosenquist, Alan Zaslavsky, Nicholas Christakis

NIH P01 AG031093

The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL ARTICLE

The Spread of Obesity in a Large Social Network Over 32 Years

Nicholas A. Christakis, M.D., Ph.D., M.P.H., and James H. Fowler, Ph.D.

Christakis and Fowler (2007): “A person’s **chances** of becoming obese increased by 57% if he or she had a friend who became obese in a given interval”

Social contagion ("peer effects") of health traits

- Several health traits considered:
  - Obesity and smoking: Christakis and Fowler (2007, 2008)
  - Happiness: Fowler and Christakis (2008)
  - Other: Alcohol use, depression ...
- Causal effect or association?
- Discussion/contributions:
  - Cohen-Cole and Fletcher (2008); Lyons (2010); Shalizi and Thomas (2011); Noel and Nyhan (2011); VanderWeele (2011); VanderWeele, Ogburn, and Tchetgen Tchetgen (2012); O'Malley (2013)

Peer Effect Confounding

- Other mechanisms leading to similarity/dissimilarity
- Homophily: "Birds of a feather flock together"
  - Hang around individuals with similar habits (e.g., over-eating, smoking); then become friends
  - Informative tie-dissolution
- Unmeasured common cause
  - Shared contextual exposure
  - Obesity/dieting propaganda, local pollutant, gym opening
  - Unobserved friends or other peers in common

Why do peer effects matter in Medicine?

- Designing behavioral interventions
  - Group interventions
  - Targeted ("seeded") interventions
- Evaluate full effect of an intervention ("collateral effects"); Spillover effects for neighborhoods (Sobel 2006)
- Interest in modeling diffusion of innovations
  - Coleman (1957): Social network/social structure related to diffusion
  - Important topic in health care policy!

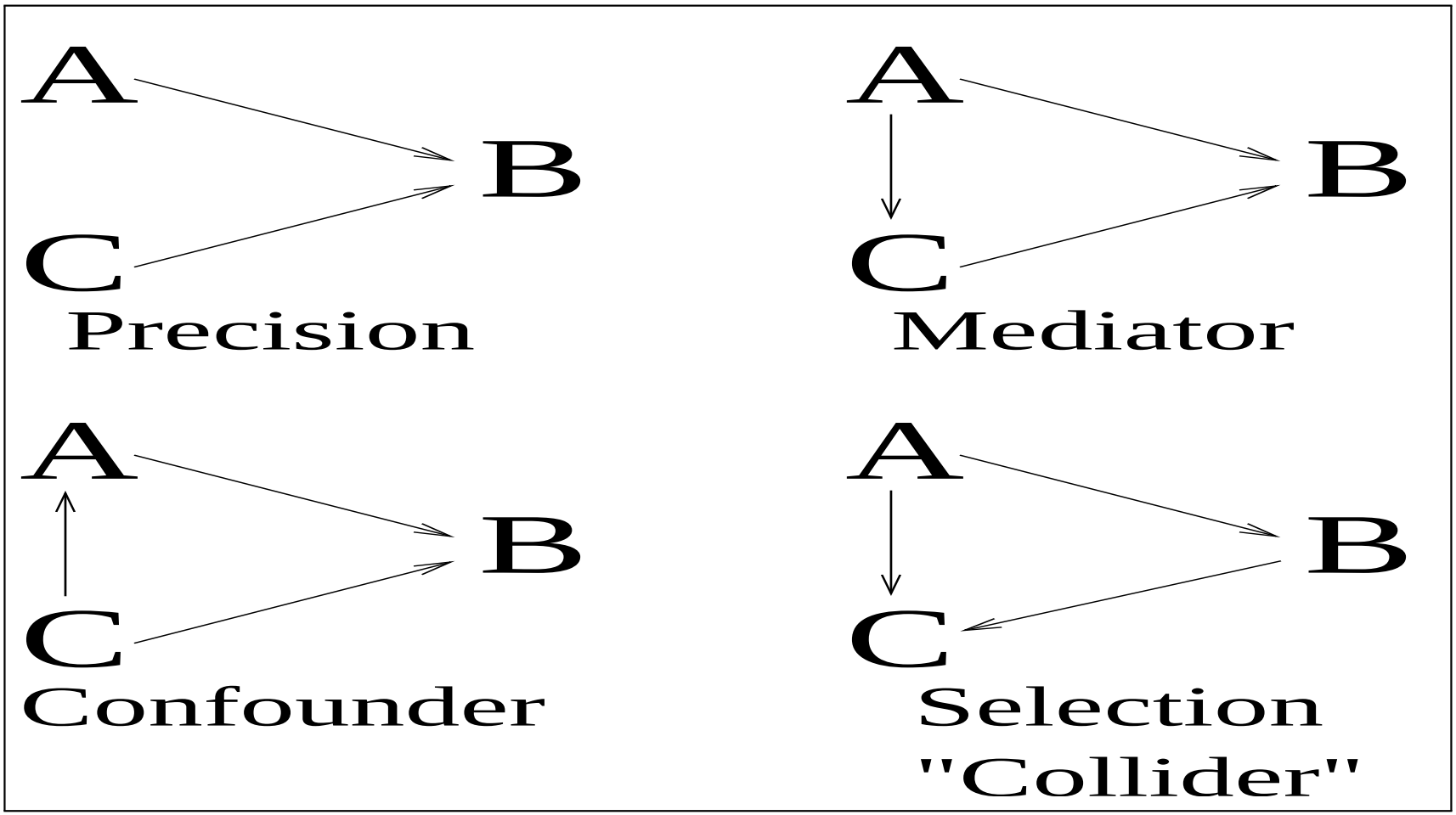
Remainder of Talk

- Formalize causal problem using Directed Acyclic Graphs (DAGs)
  - Example phenotype: Body Mass Index (BMI)
- Describe Instrumental variables (IV)-based solutions
  - Example: genetic alleles
  - Application: Framingham Heart Study
- Results
- Discussion

Directed Acyclic Graphs: Basics

- Pearl (1995, 2009); Elwert (2013)
- Nonparametric depiction of causal dependencies
- Nodes: Variables
- Edges: Direct causal effects in direction of arrow
- Missing edges: Strictly zero effect ("exclusion restrictions")
  - Helpful for identification
- DAG $\neq$ graph of a network!

Simple DAGs: Effect of A on B in presence of C

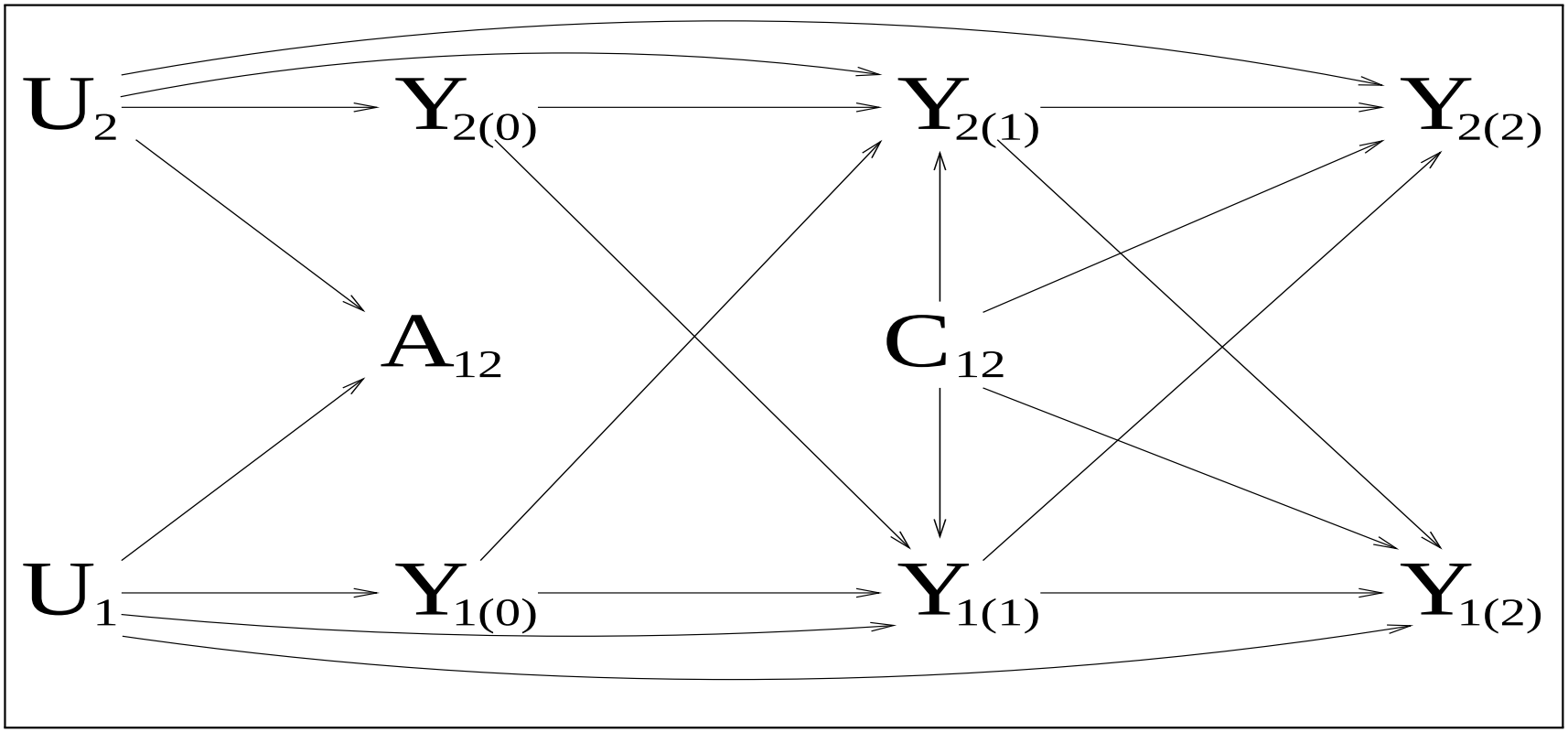


Need longitudinal data to discern role of C

Technical details

- DAGs translate between causal assumptions and observable associations
- A path is d-separated or blocked if:
  - It contains a non-collider variable that has been conditioned on, or
  - If it contains a collider variable and neither the collider nor any of its descendants has been conditioned on.
- Variables that are d-separated along all paths are statistically independent (Balke & Pearl 1988). Otherwise, they may be associated.

DAG illustrating homophily and common cause



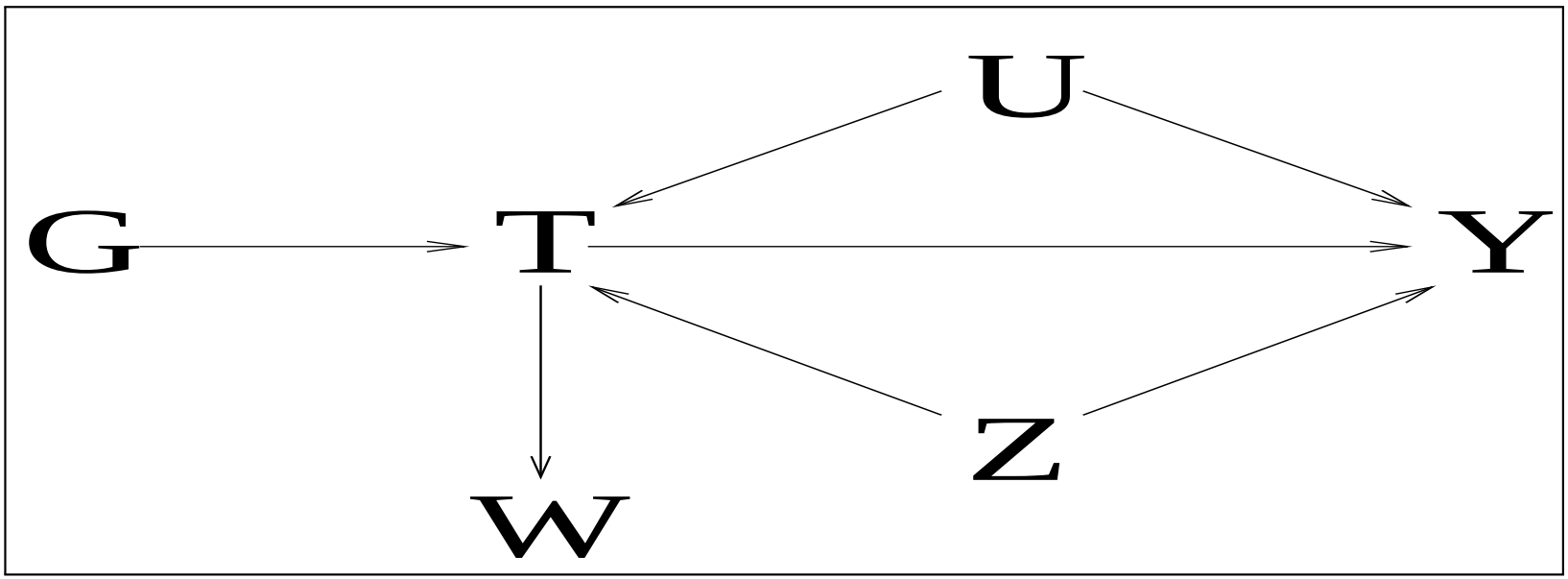
Dyad of individuals 1 and 2 observed for $\boldsymbol{t}$*=(0), (1), (2) following tie-formation.* $\boldsymbol{Y}$ *= BMI,* $\boldsymbol{A}$ *= relationship status,* $\boldsymbol{U}$ *= unobserved individual predictors,* $\boldsymbol{C}$ *= unobserved "shared" exposure*

The Challenge

- Association of $\boldsymbol{Y_{2(t-1)}}$ ("treatment") and $\boldsymbol{Y_{1t}}$ ("outcome") includes:
  - Common cause confounding (Christakis and Fowler, 2007): $\boldsymbol{Y_{2(t-1)} \leftarrow C_{12} \rightarrow Y_{1t}}$
  - Homophily (Shalizi and Thomas, 2011): $\boldsymbol{Y_{2(t-1)} \leftarrow U_2 \rightarrow [A_{12}] \leftarrow U_1 \rightarrow Y_{1t}}$
- Longitudinal data helps in special cases:
  - For example, temporary effects of $\boldsymbol{U, C}$
- We take less parametric approach to Steglich et al. (2010)

Definition of an Instrumental Variable (IV)

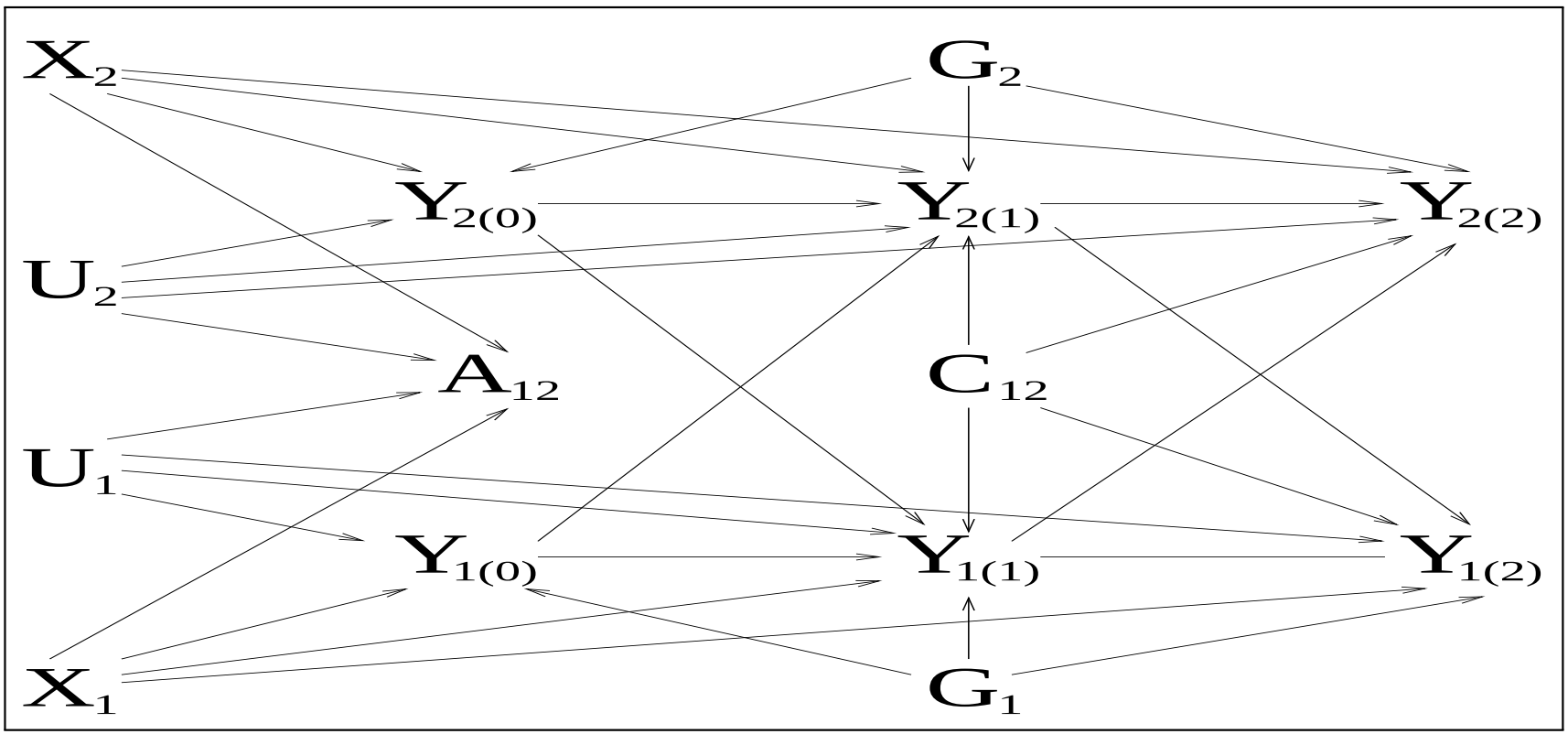
- Brito and Pearl (2002)
- There exists an unblocked path from the IV, G, to the treatment, T
- There are no unblocked paths from G to the outcome, Y, other than through T
- Can condition on observed covariates, denoted Z
  - But cannot condition on W, a descendant of T



Genetic Alleles: IVs for peer effects?

- Genetic alleles fixed at conception
  - Measurable any time
  - Endure over lifetime
- Genes linked to obesity and BMI:
  - Fat mass and obesity gene (FTO)
  - Melanocortin-4 receptor gene (MC4R)
- Homozygous (risk and non-risk) and heterozygous states

DAG with Genes as IVs



Dyad of individuals 1 and 2 observed for $\boldsymbol{t}$*=(0), (1), (2) following tie-formation.* X *= observed predictors;* Z *may include* $\{\mathbf{X}, \mathbf{G}_1, \mathbf{A}\}$

Requirements for Genes as IVs

- Need to account for backdoor path through $\boldsymbol{Y_{2(0)}}$
- But conditioning on $\boldsymbol{Y_{2(0)}}$ opens backdoor path through $\boldsymbol{U_2}$
- Solution: instrument both $\boldsymbol{Y_{2(1)}}$ and $\boldsymbol{Y_{2(0)}}$
- Each treatment must have its own unblocked path from the IV
  - Therefore, $\boldsymbol{\dim(G_2) \geq 2}$
- Elements of $\boldsymbol{G_2}$ do not need to be valid IVs individually!

Justification: IV Set Criterion (Brito and Pearl, 2002)

Let $\boldsymbol{T} = (\boldsymbol{T_1}, \ldots, \boldsymbol{T_L})$ be a multivariate treatment and $\boldsymbol{\mathcal{D}_{\text{test}}}$ be $\boldsymbol{\mathcal{D}}$ after removing all edges emanating from $\boldsymbol{T}$. Then $\boldsymbol{G} = (\boldsymbol{G_1}, \ldots, \boldsymbol{G_K})$ is an IV-set for the joint causal effect of $\boldsymbol{T}$ on $\boldsymbol{Y}$ conditional on a set of variables $\boldsymbol{Z}$ if:

1. $\boldsymbol{Z}$ contains no descendant of $\boldsymbol{T}$ in $\boldsymbol{\mathcal{D}}$.
2. For every $\boldsymbol{l} \in \{\boldsymbol{1}, \ldots, \boldsymbol{L}\}$ there exists an unblocked path, called $\boldsymbol{\text{path}_l}$, between $\boldsymbol{G_k} \in \boldsymbol{G}$ and $\boldsymbol{T_l} \in \boldsymbol{T}$ in $\boldsymbol{\mathcal{D}_{\text{test}}}$ after conditioning on $\boldsymbol{Z}$, such that $\{\boldsymbol{\text{path}_1}, \ldots, \boldsymbol{\text{path}_L}\}$ have no nodes in common.
3. For $\boldsymbol{k} \in \{\boldsymbol{1}, \ldots, \boldsymbol{K}\}$ there are no unblocked paths between $\boldsymbol{G_k} \in \boldsymbol{G}$ and $\boldsymbol{Y}$ in $\boldsymbol{\mathcal{D}_{\text{test}}}$ after conditioning on $\boldsymbol{Z}$.

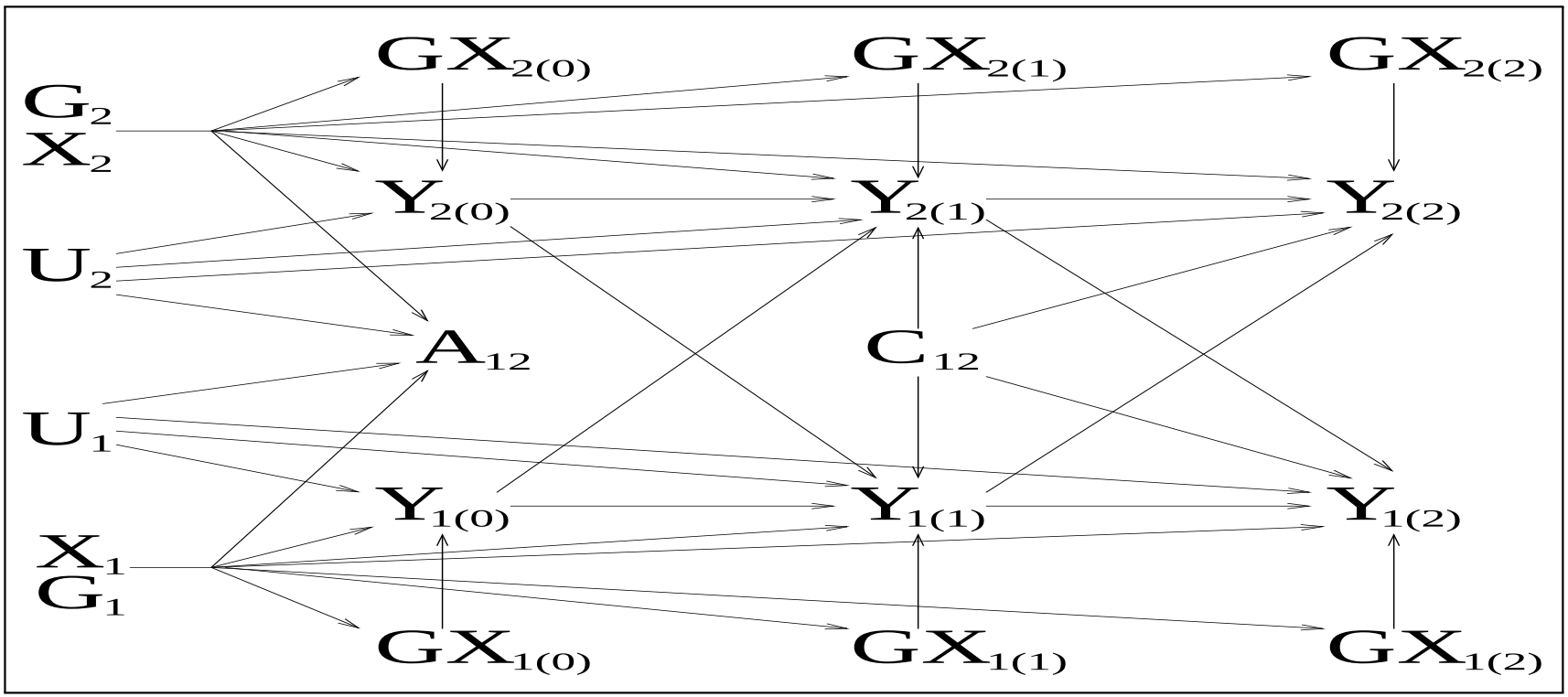
What if gene effects span four, five ... periods

- Instrument all lagged treatment variables back to tie-formation: $\boldsymbol{Y_{j(t-1)}, \ldots, Y_{j(0)}}$
- Might be okay if $\boldsymbol{\dim(G_2)}$ sufficiently big
- In general, becomes impractical

Using gene-expression as the IV

- Gene expression changes with age?
  - Let $\boldsymbol{GX_{2t}} = \boldsymbol{G_2} \times \mathbf{Age_{2t}}$
- Age is a likely modifier of gene expression
  - Hypothesis: Relative influence of genetic state varies with age
  - Allow different effects of genetic states with age
- Confounders must have same form of dependence (linear, quadratic ...) by age to create problems

DAG with time-varying IV



Dyad of individuals 1 and 2 observed for $\boldsymbol{t}$*=(0), (1), (2) following tie-formation.* $\boldsymbol{Y}$ *= BMI,* $\boldsymbol{A}$ *= relationship status,* $\boldsymbol{U}$ *= Unobserved individual predictors,* $\boldsymbol{C}$ *= unobserved "shared" predictors,* $\boldsymbol{G}$ *= Genes,* $\boldsymbol{X}$ *= Observed predictors,* $\boldsymbol{GX}$ *= Gene-expression (by age)*

Two-stage least squares (2SLS) procedure

- Let $\boldsymbol{Z}$ = conditioning set; e.g., $\boldsymbol{X}, \boldsymbol{G_1}, \boldsymbol{A}, \{\boldsymbol{Y_{(t-l)}}\}_{t \geq l \geq 1}$
- Stage I:

$$\boldsymbol{Y_{2(t-1)}} = \boldsymbol{G X^T_{2(t-1)} \theta_1} + \boldsymbol{Z^T_{(t)} \theta_2} + \boldsymbol{\delta_{1(t)}},$$

- Compute fitted values: $\boldsymbol{\hat{Y}_{2(t-1)}}$
- Stage II:

$$\boldsymbol{Y_{1(t)}} = \boldsymbol{\alpha_1 \hat{Y}_{2(t-1)}} + \boldsymbol{Z^T_{(t)} \beta} + \boldsymbol{\hat{\epsilon}_{1(t)}},$$

where $\boldsymbol{\hat{\epsilon}_{1(t)}} = \boldsymbol{\epsilon_{1(t)}} + \boldsymbol{\alpha_1 (Y_{2(t-1)} - \hat{Y}_{2(t-1)})}$.

Specification of Z

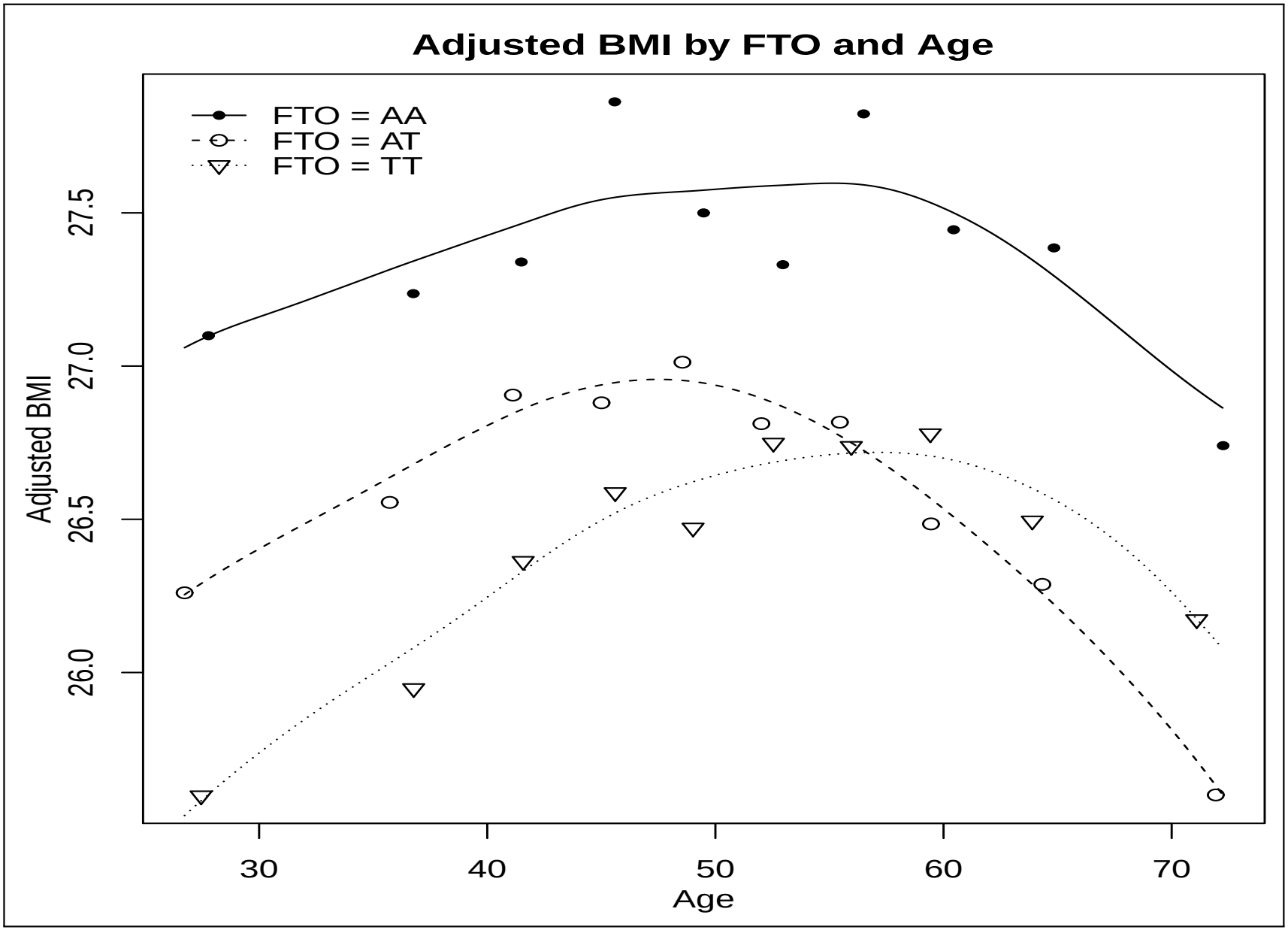
- Include ego fixed effects.
  - Block time invariant confounders
  - Account for (time-invariant) population stratification
- Include predictors in Z to account for pleiotropy
- Various complications accommodated:
  - Phenotypic homophily (at $t = (0)$)
  - Genetic homophily

Application: Framingham heart study (FHS) network

- Offspring cohort: 5,124 individuals; 7 medical exams (1971-2003)
  - Slightly modified dataset to what Christakis and Fowler et al analyzed.
- Two types of relationships (analyze separately)
  - A "close friend" (named friend at exam)
  - Spouse

Key individual variables

- Outcome: Body mass index (BMI), $\boldsymbol{Y_{1t}}$, $t = 1, \ldots, 7$
- Key predictor, BMI of alter, $\boldsymbol{Y_{2(t-1)}}$
- Discretionary exogeneous predictors, $\boldsymbol{X_{1t}}$:
  - Gender, age, gender $\times$ age, birth-year, birth-era, location
  - Smoker, married, number of siblings, health, SES
- Recall IVs, $\boldsymbol{GX_{2(t-1)}}$
  - Fat mass and obesity gene (FTO)
  - Melanocortin-4 receptor gene (MC4R) interacted with age


Adjusted BMI by FTO and Age
FTO = AA
FTO = AT
FTO = TT
Adjusted BMI
26.0
26.5
27.0
27.5
30
40
50
60
70
Age

Estimated peer effects under four specifications using 2SLS

$GX_{2(t-2)}$	$Y_{1(t-1)}$	F_5	Estimate	95% CI	
Ego nominated friend					
Exclude	Exclude	2.150	**0.888**	**(0.063,**	**1.713)**
Exclude	Covariate	1.731	0.874	(-0.031,	1.779)
Covariate	Exclude	1.181	0.133	(-0.796,	1.062)
Covariate	Covariate	1.144	-0.003	(-0.911,	0.906)
Spouse					
Exclude	Exclude	4.064	0.099	(-0.324,	0.522)
Exclude	Covariate	4.351	0.101	(-0.287,	0.488)
Covariate	Exclude	0.268	-0.102	(-1.855,	1.652)
Covariate	Covariate	0.181	0.906	(-1.832,	3.643)

OLS effects non-significant for friends but significant for spouses (estimates approx 0.05). F_5 *is statistic of stage I F-test*

Conclusion

- Time-invariant IVs difficult to justify
- Gene-expression provides viable approach
  - Attempts to account for all sources of confounding
  - But results in weak identification
- Could use alter covariates that **do not** influence relationship status as additional IVs (Haining, 1978; Kelejian and Robinson 1993)
  - Contextual variables in spatial econometrics
  - May strengthen identification
- Exploratory idea: use auxillary data to build gene-phenotype model
  - Different structure than stage I equation
  - Project multiple weak IVs to a strong scalar IV

Future work

- Include mediator variables (e.g., nutritional habits)
- Sociocentric analyses
- Account for evolution of relationships (O'Malley and Christakis, 2011 SIM; Paul and O'Malley, 2013 JRSS-C)

Extensions of basic methodology to account for heterogeneous peer effects

- Can extend approach to account for peer effect heterogeneity
  - Meaning or strength of relationship might be directional
  - Each dyad contains two "experiments"
  - $\boldsymbol{Y_{2(t-1)} \rightarrow Y_{1t}}$ is one, $\boldsymbol{Y_{1(t-1)} \rightarrow Y_{2t}}$ is another.
- Account for individuals in multiple dyads
  - Extreme case: complete sociocentric data
- Account/utilize variation in tie-status

References

- O'Malley AJ, Marsden PV. (2008). The Analysis of Social Networks *Health Services & Outcomes Research Methodology*, 8, 222–269
- O'Malley AJ, Christakis NA. (2011). Longitudinal Analysis of Large Social Networks: estimating the Effect of Health Traits on changes in Friendship Ties. *Statistics in Medicine*, 30, 9, 950-964
- OMalley AJ. (2013). The Analysis of Social Network Data: An Exciting Frontier for Statisticians. Statistics in Medicine, 32, 539-555
- Christakis NA, Fowler JH. (2007). The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine*, 357, 370–379
- Christakis NA, Fowler, JH (2008). Dynamics of Smoking Behavior in a Large Social Network. *New England Journal of Medicine*, 358, 2249–2258

References (cont.)

- Fowler JH, Christakis NA. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *New England Journal of Medicine*, 337, doi:10.1136/bmj.a2338.
- Lyons R. (2010). The Spread of Evidence-Poor Medicine via Flawed Social-Network Analysis. *Statistics, Politics, and Policy*, 2, Article 2, doi:10.2202/2151-7509.1024.
- Cohen-Cole E, Fletcher JM. (2008). Detecting Implausible Social Network Effects in Acne, Height, and Headaches: Longitudinal Analyses. *British Medical Journal*, 337, a2533.
- Shalizi CR, Thomas AC. (2011). Homophily and Contagion Are Generically Confounded in Observational Social Network Studies, *Sociological Methods & Research*, 40, 211–239.

References (cont.)

- Land KC, Deane G. (1992). On the Large-Sample Estimation of Regression Models with Spatial or Network Effect Terms: A Two-Stage Least-Squares Approach. *Sociological Methodology*, (ed: Peter V. Marsden), Oxford, UK: Basil Blackwell, Ltd, 221–248
- Anselin L. (1988). *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers: Dordrecht, The Netherlands
- Sargan JD. (1958). The Estimation of Econometric Relationships Using Instrumental Variables, *Econometrica*, 26, 393–415

References (cont.)

- Haining, R.P. (1978). The moving average model for spatial interaction. Transactions of the Institute of British Geographers, 3, 202-225.
- Kelejian, H.H., Robinson, D.P., 1993. A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model. Papers in Regional Science 72, 297-312.