Homophily, Contagion, Confounding: Pick Any Three

Cosma Shalizi

Statistics Department, Carnegie Mellon University

Santa Fe Institute

11 December 2009

My interest: non-parametric reconstruction of dynamical systems from the behavior they generate

Perspective: Yet another ex-physicist

$\therefore$ Social networks are "just" large coupled dynamical systems

Apologies in advance for social-scientific and graphological naivete

“If your friend Joey jumped off a bridge, would you jump too?”

"If your friend Joey jumped off a bridge, would you jump too?"

1. yes: Joey inspires you (social contagion or influence)

"If your friend Joey jumped off a bridge, would you jump too?"

1. yes: Joey inspires you (social contagion or influence)
2. yes: Joey infects you with a parasite which suppresses fear of falling (actual contagion)

"If your friend Joey jumped off a bridge, would you jump too?"

1. yes: Joey inspires you (social contagion or influence)
2. yes: Joey infects you with a parasite which suppresses fear of falling (actual contagion)
3. yes: you're friends *because* you both like to jump off bridges (manifest homophily)

"If your friend Joey jumped off a bridge, would you jump too?"

1. yes: Joey inspires you (social contagion or influence)
2. yes: Joey infects you with a parasite which suppresses fear of falling (actual contagion)
3. yes: you're friends *because* you both like to jump off bridges (manifest homophily)
4. yes: you're friends *because* you both like roller-coasters, and have a common risk-seeking propensity (latent homophily)

"If your friend Joey jumped off a bridge, would you jump too?"

1. yes: Joey inspires you (social contagion or influence)
2. yes: Joey infects you with a parasite which suppresses fear of falling (actual contagion)
3. yes: you're friends *because* you both like to jump off bridges (manifest homophily)
4. yes: you're friends *because* you both like roller-coasters, and have a common risk-seeking propensity (latent homophily)
5. yes: because you're both on it when it starts collapsing and that's the only way off (external causation)

Wikipedia, s.v. “Tacoma Narrows Bridge (1940)”

Are these distinctions with *observational* differences?

Are these distinctions with *observational* differences?

1. Can't experiment by pushing Joey off the bridge

Are these distinctions with *observational* differences?

1. Can't experiment by pushing Joey off the bridge
2. Don't want to impose strong parametric assumptions

Are these distinctions with *observational* differences?

1. Can't experiment by pushing Joey off the bridge
2. Don't want to impose strong parametric assumptions

Manski (1993) suggests this is just not identifiable, but does not quite settle the problem

Influence due to group average vs. individuals

Contagion, Influence

Whether i does something at time t is well-predicted by whether i's neighbors had already done it at $t-1$

- Diffusion of innovations
- Infectious diseases
- Not-obviously-infectious conditions (e.g., obesity) ...

This *can* be due to influence or contagion

Analogy of ideas to diseases is very old: Pliny used it in 110 (*Epistles* X 96)

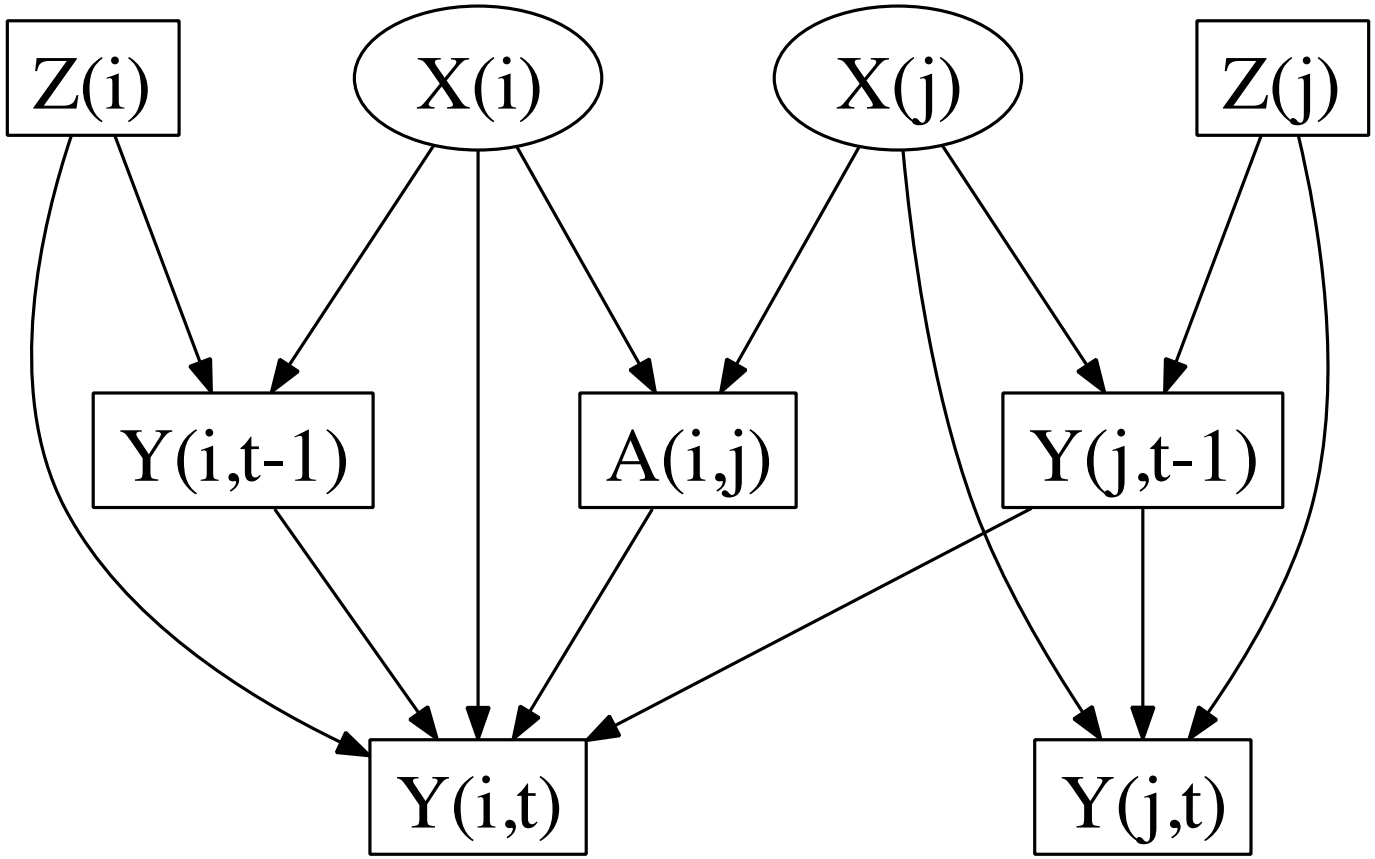
Can the same *observational* consequences can follow from latent homophily?

Notation:

- $Y(i,t) =$ does node i show condition/behavior at time t?
- $X(i) =$ *latent* persistent trait of i
- $Z(i) =$ other, manifest persistent traits
- $A(i,j) =$ whether there is an edge from j to i

We suppose that:

- $Y(i,t-1)$ has a direct influence on $Y(i,t)$
- $X(i)$ has a direct influence on whether/when i adopts
- $Z(i)$ has a direct influence on $Y(i,t)$ (possibly null)
- $Y(j,t-1)$ *may* have a direct influence on $Y(i,t)$, but only if $A(i,j) = 1$
- Homophily: $X(i)$ and $X(j)$ both directly influence $A(i,j)$


Z(i)
X(i)
X(j)
Z(j)
Y(i,t-1)
A(i,j)
Y(j,t-1)
Y(i,t)
Y(j,t)

Contagion Effects are Nonparametrically Unidentifiable

Informally:

1. $Y(j, t-1)$ is informative about $X(j)$
2. $X(j)$ is informative about $X(i)$ if i and j are neighbors
3. $X(i)$ is informative about $Y(i, t)$
4. $\therefore Y(i, t) \not\perp\!\!\!\perp Y(j, t-1)$, even if there is no direct causal effect
5. $\therefore$ Latent homophily is confounded with contagion

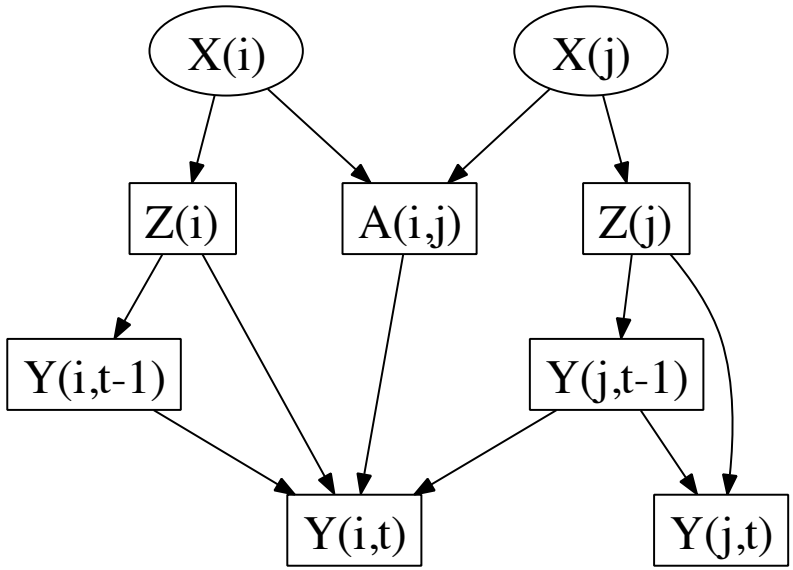
More formally:

1. $Y(i,t) \leftarrow X(i) \rightarrow A(i,j)$ is a confounding path from $Y(i,t)$ to $A(i,j)$
2. Likewise $Y(j,t-1) \leftarrow X(j) \rightarrow A(i,j)$ is a confounding path from $Y(j,t-1)$ to $A(i,j)$
3. $\therefore$ the direct effect of $Y(j,t-1)$ on $Y(i,t)$ is not identifiable (Pearl, 2009, §3.5, pp. 93–94)

Adding conditioning on $Y(i,t-1)$ and $Y(j,t)$ does not remove the confounding paths
Neither does adding conditioning on $Z(i), Z(j)$
Argument still goes through with time-varying edges (more spaghetti)

Getting Identifiability

Parametric assumptions *can* suffice
Better: condition on X; or find Z which block paths from Y to X
Explicit modeling as in Leenders (1995); Steglich *et al.* (2004) does both



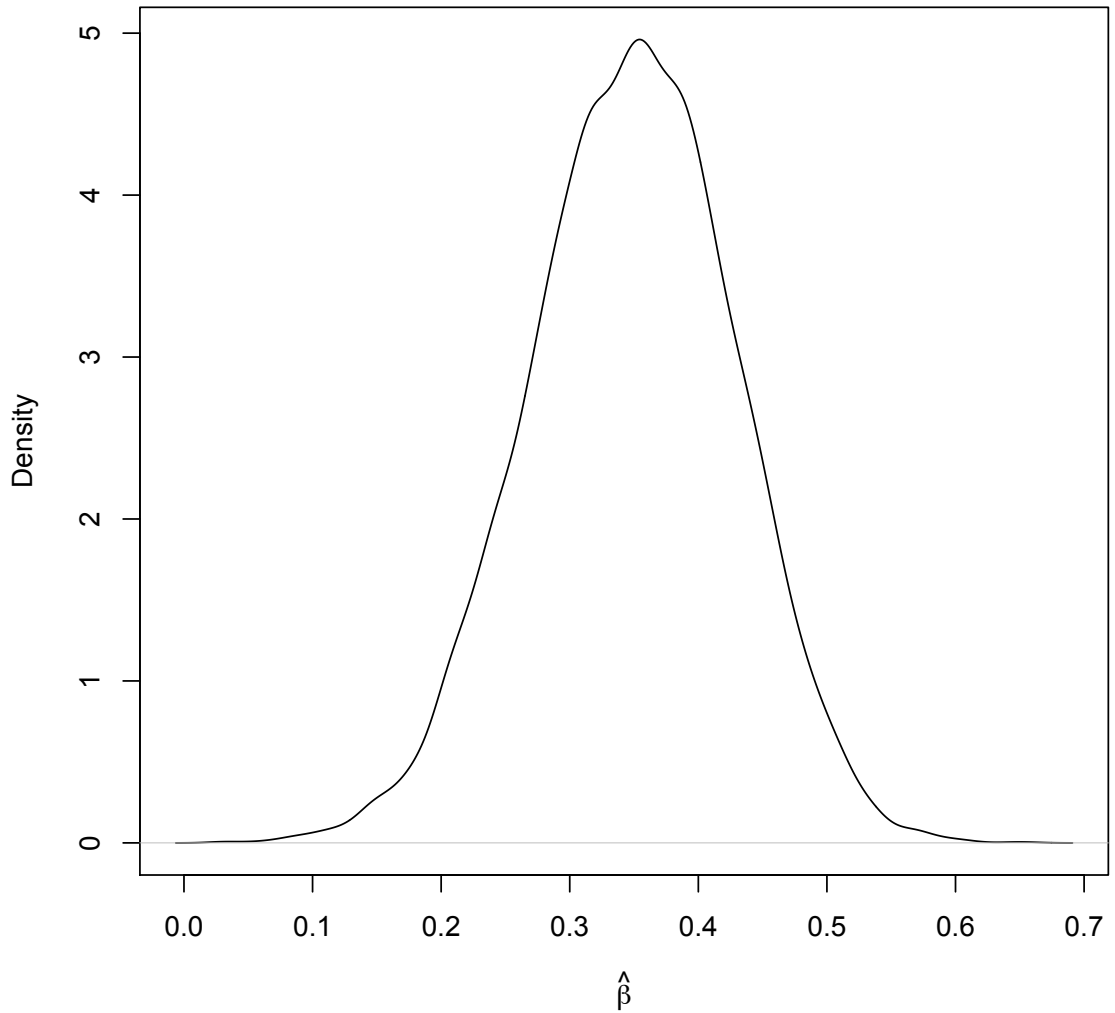
The Argument from Asymmetry

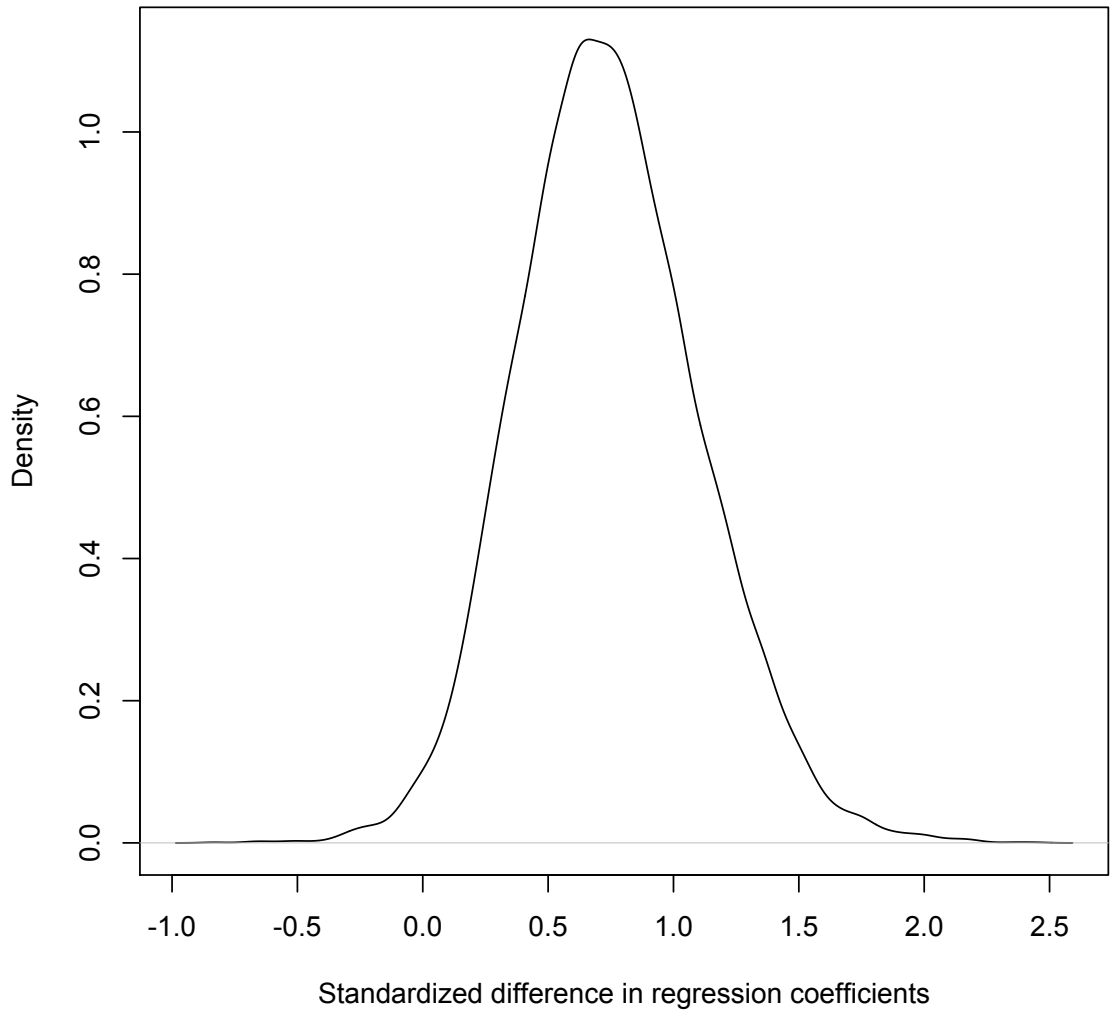
Focus on unreciprocated edges, $i \rightarrow j$, $j \not\rightarrow i$
Suppose $Y(i,t)|Y(j,t-1)) \not\sim Y(j,t)|Y(i,t-1)$
Doesn't this argue for direct influence?
Considerable *prima facie plausibility*
Argument breaks down if senders and receivers have systematically different values of X, with different local relations to Y

Toy Example

Ignore time-dependence; try to predict $Y(i)$ from $Y(j)$ and vice versa when $A_{ij} = 1, A_{ji} = 0$
$X(i) \sim \mathcal{U}(0,1)$
Edges form with probability $\propto \text{logit}^{-1}(-3|X(i) - X(j)|)$
i nominates j from among neighbors, $\propto \text{logit}^{-1}(-|X(j) - 0.5|)$
$Y(i) = 10(X(i) - 0.5)^3 + \mathcal{N}(0, 0.1)$
Results:

- $Y(i)$ is well-predicted from $Y(j)$
- *Nominees* are disproportionately in the middle; $i \to j, j \not\to i$ suggests i is more peripheral
- For asymmetric pairs, regression of sender on receiver differs from that of receiver on sender


Sampling distribution: regression coefficient of Y(i) on Y(j)
Density
0
1
2
3
4
5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
$\hat{\beta}$


Asymmetry from preferential nomination
Density
0.0
0.2
0.4
0.6
0.8
1.0
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Standardized difference in regression coefficients

Making homophily and contagion look like causation

Long-term, hard-to-change social/economic status explains more short-term, malleable cultural / political / consumer variables

Making homophily and contagion look like causation

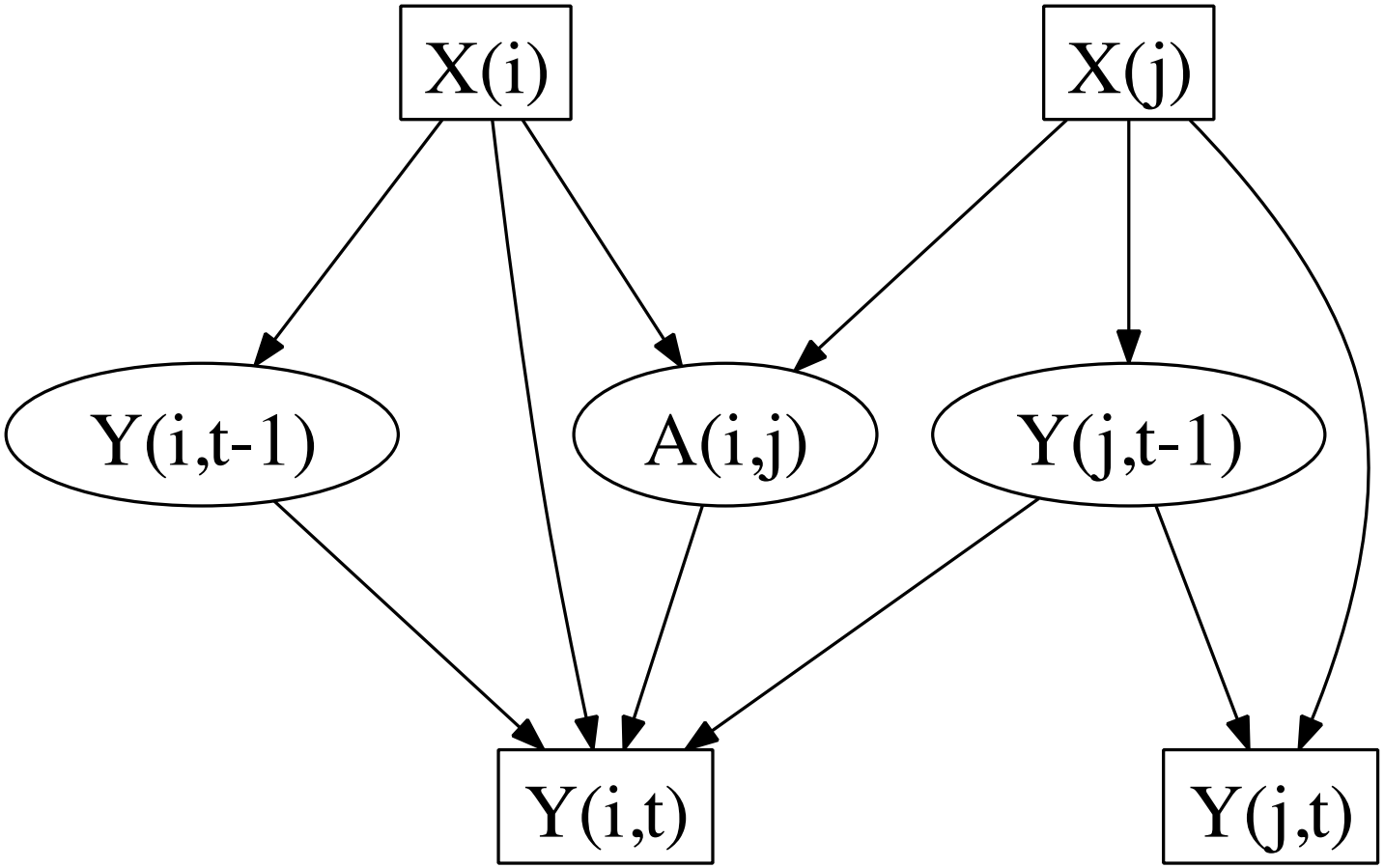
Long-term, hard-to-change social/economic status explains more short-term, malleable cultural / political / consumer variables

> *Gellner: "Social structure is who you can marry, culture is what you wear at the wedding."*

What's the evidence?

- The stories sound good
- Casual empiricism
- Correlation/regression analyses; cultural choices are predictable from social positions (e.g. Bourdieu (1984))

Probably even true a lot of the time
BUT usually ignores social networks and just looks at surveys


X(i)
X(j)
Y(i,t-1)
A(i,j)
Y(j,t-1)
Y(i,t)
Y(j,t)

More Confounding

Direct influence of $X(i)$ on $Y(i,t)$ is confounded with contagion:

1. $X(i)$ is a cue about who i's friends are, i.e. $A(i,j)$
2. $\therefore X(i)$ is a cue about what i's friends think, $Y(j,t-1)$
3. contagion: $Y(j,t-1)$ influences $Y(i,t)$ if $A(i,j)=1$
4. $\therefore X(i) \not\perp Y(i,t)$ even if no direct influence

Responsible Just-So Story-telling

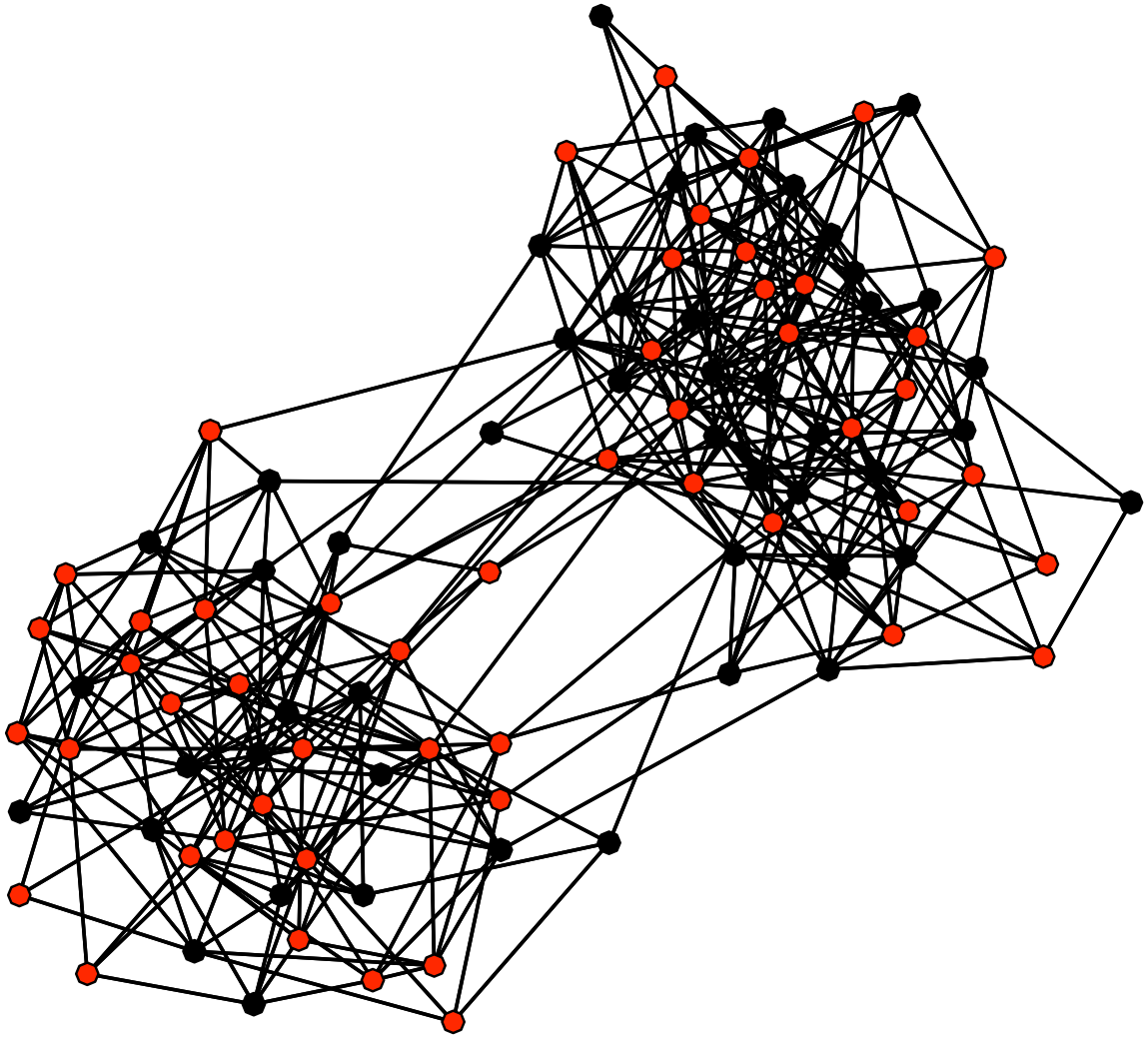
These accounts are usually adaptationist/functionalist
At the very least they are causal accounts
We should really check them
Biology suggests: a **neutral model**

- Include all the evolutionary processes *except* adaptation
- Work out expected behavior of this model
- Data departing from neutral model $\Rightarrow$ evidence of adapation

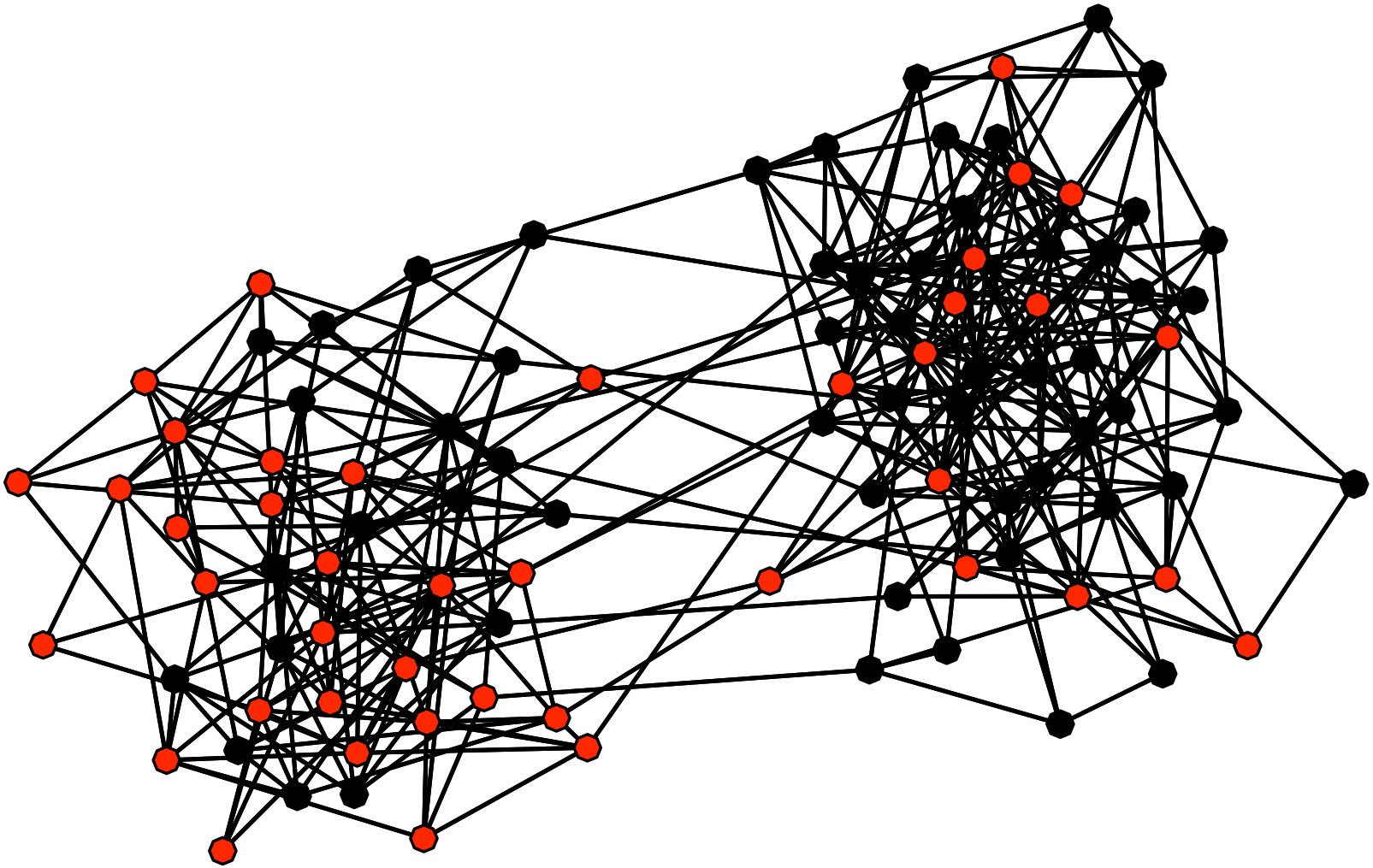
Caricature Neutral Model of Cultural Evolution

- $X(i) =$ unchanging status variable for node i ("social")
- Network is assortative on X (minimal departure from Erdős-Rényi)
- $Y(i,t) =$ rapidly changing choice variable for i ("cultural")
  1. At each t, pick a random i, and a random neighbor j
  2. Set $Y(i,t) = Y(j,t-1)$
  3. Go to (1)
- $Y(\cdot, 0) =$ Bernoulli(1/2) process

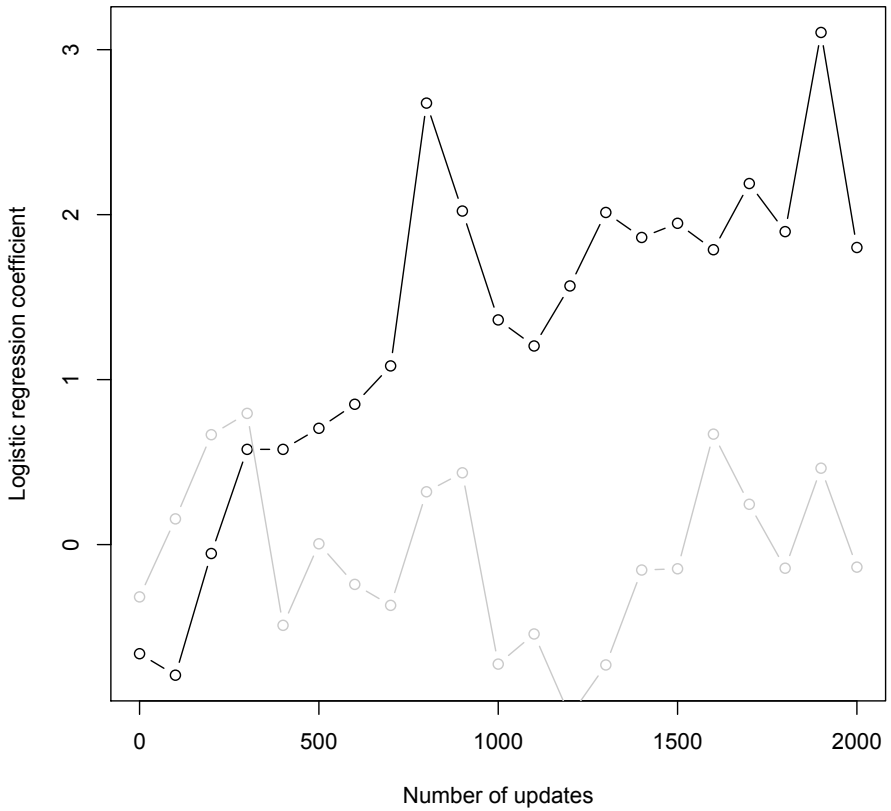
(= "voter model" of statistical mechanics)



100 node network, homophily for status (2 groups), initial choices



After 1000 updates



100 node network

Logistic regression of choice on status
black = assortative
grey = non-assortative

- Neutral diffusion + homophily looks like a real connection between social status and cultural choices
- Problem is *not* the ecological fallacy (red-state/blue-state fallacy) (not using aggregated data)
- Problem is that choices are not independent conditional on statuses
- Deconfound by conditioning on previous Y_j of neighbors

Partial Control by Clustering?

If the problem is latent heterogeneity, why not try to identify the latent trait?
Latent homophily $\Rightarrow$ you tend to resemble your neighbors
$\Rightarrow$ Especially likely if you all have lots of neighbors in common who all have lots of neighbors in common, etc.
$\Rightarrow$ modules/communities
Can't remove confounding but *might* reduce it
. . . or make it worse if the latent relationship isn't simple homophily (e.g. block models)

An Analogy For Community Control

Gene association studies: does having this genetic variant influence this trait/change this risk?

Real populations are structured

Sub-populations differ (due to reproductive isolation etc.)

$\Rightarrow$ genes are correlated

$\Rightarrow$ random biases and inflated variances (vs. usual formulas)

$\Rightarrow$ many bogus results

Population structure substantial even for e.g. Germany (Steffens *et al.*, 2006) or Italy, never mind "white Americans"

Responses: (1) pedigrees; (2) "genomic control" by estimating over-dispersion empirically (Devlin *et al.*, 2001); (3) clustering — the diffusion maps in Lee *et al.* (2009) look *a lot* like Newman (2006)

Conclusion

1. Homophily + causal influence looks like contagion
2. Homophily + contagion looks like causal influence
3. Of course contagion + causality looks like (is?) homophily

Bourdieu, Pierre (1984). *Distinction: A Social Critique of the Judgement of Taste*. Cambridge, Massachusetts: Harvard University Press.

Devlin, B., Kathryn Roeder and Larry Wasserman (2001). "Genomic Control, a New Approach to Genetic-Based Association Studies." *Theoretical Population Biology*, **60**: 155–166. URL `http://www.stat.cmu.edu/tr/tr749/tr749.html`. doi:10.1006/tpbi.2001.1542.

Lee, Ann B., Diana Luca, Lambertus Klei, Bernie Devlin and Kathryn Roeder (2009). "Discovering Genetic Ancestry Using Spectral Graph Theory." *Genetic Epidemiology*, **34**: 51–59. URL `http://www.stat.cmu.edu/~roeder/pubs/llkdr2009.pdf`. doi:10.1002/gepi.20434.

Leenders, Roger Th. A. J. (1995). *Structure and Influence:*

Statistical Models for the Dynamics of Actor Attributes, Network Structure and Their Interdependence. Amsterdam: Thesis Publishers.

Manski, Charles F. (1993). "Identification of Endogeneous Social Effects: The Reflection Problem." *Review of Economic Studies*, **60**: 531–542.

Newman, Mark E. J. (2006). "Finding Community Structure in Networks Using the Eigenvectors of Matrices." *Physical Review E*, **74**: 036104. URL `http://arxiv.org/abs/physics/0605087`.

Pearl, Judea (2009). *Causality: Models, Reasoning, and Inference*. Cambridge, England: Cambridge University Press, 2nd edn.

Steffens, Michael, Claudia Lamina, Thomas Illig, Thomas Bettecken, Rainer Vogler, Patricia Entz, Eun-Kyung Suk,

Mohammad Reza Toliat, Norman Klopp, Amke Caliebe, Inke R. König, Karola Köhler, Jan Lüdemann, Amalia Diaz Lacava, Rolf Fimmers, Peter Lichtner, Andreas Ziegler, Andreas Wolf, Michael Krawczak, Peter Nürnberg, Jochen Hampe, Stefan Schreiber, Thomas Meitinger, H.-Erich Wichmann, Kathryn Roeder, Thomas F. Wienker and Max P. Baur (2006). "SNP-Based Analysis of Genetic Substructure in the German Population." *Human Heredity*, **62**: 20–29. doi:10.1159/000095850.

Steglich, Christian, Tom A. B. Snijders and Michael Pearson (2004). *Dynamic Networks and Behavior: Separating Selection from Influence*. Tech. Rep. 95-2001, Interuniversity Center for Social Science Theory and Methodology, University of Groningen. URL `http://www.stats.ox.ac.uk/~snijders/siena/`

SteglichSnijdersPearson2009.pdf.