Homophily, Contagion, Confounding: Pick Any Three

Cosma Shalizi

Statistics Department, Carnegie Mellon University

Santa Fe Institute

19 May 2011
Atlantic Causal Inference Conference

My real interest: how much of the mechanism of a complex system can we reconstruct from observations?

My real interest: how much of the mechanism of a complex system can we reconstruct from observations?
As Yet Another Ex-Physicist, social networks are just large coupled dynamical assemblage...

My real interest: how much of the mechanism of a complex system can we reconstruct from observations?

As Yet Another Ex-Physicist, social networks are just large coupled dynamical assemblage. . .

Got interested in this problem with Sperber (1996) and especially Christakis and Fowler (2007)

My real interest: how much of the mechanism of a complex system can we reconstruct from observations?

As Yet Another Ex-Physicist, social networks are just large coupled dynamical assemblage. . .

Got interested in this problem with Sperber (1996) and especially Christakis and Fowler (2007)

. . . and expected to reach much more positive conclusions

My real interest: how much of the mechanism of a complex system can we reconstruct from observations?

As Yet Another Ex-Physicist, social networks are just large coupled dynamical assemblage. . .

Got interested in this problem with Sperber (1996) and especially Christakis and Fowler (2007)

. . . and expected to reach much more positive conclusions

Apologies in advance for social-scientific and etiological naivete

My real interest: how much of the mechanism of a complex system can we reconstruct from observations?

As Yet Another Ex-Physicist, social networks are just large coupled dynamical assemblage. . .

Got interested in this problem with Sperber (1996) and especially Christakis and Fowler (2007)

. . . and expected to reach much more positive conclusions

Apologies in advance for social-scientific and etiological naivete

Joint work with Andrew Thomas, CMU Statistics

Details: Shalizi and Thomas (2011)

"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 (biological 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 (biological 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 (biological 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 (biological 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 sometimes jumping off a bridge is the only sane thing to do (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. Can't experiment by keeping Joey and Irene apart, or pushing them together

Are these distinctions with *observational* differences?

1. Can't experiment by pushing Joey off the bridge
2. Can't experiment by keeping Joey and Irene apart, or pushing them together
3. 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. Can't experiment by keeping Joey and Irene apart, or pushing them together
3. 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 some does something at one time can be predicted from whether their neighbors have already done

Contagion, Influence

Whether some does something at one time can be predicted from whether their neighbors have already done

- Infectious disease

Contagion, Influence

Whether some does something at one time can be predicted from whether their neighbors have already done

- Infectious disease
- Diffusion of innovations
- Diffusion of ideologies

Pliny (+110): Christianity is a “contagious superstition”

Contagion, Influence

Whether some does something at one time can be predicted from whether their neighbors have already done

- Infectious disease
- Diffusion of innovations
- Diffusion of ideologies

  Pliny (+110): Christianity is a “contagious superstition”
- Not-obviously-infectious conditions (e.g., obesity, loneliness, divorce) ...

Contagion, Influence

Whether some does something at one time can be predicted from whether their neighbors have already done

- Infectious disease
- Diffusion of innovations
- Diffusion of ideologies

  Pliny (+110): Christianity is a "contagious superstition"
- Not-obviously-infectious conditions (e.g., obesity, loneliness, divorce) . . .

This *can* be due to influence or contagion

Contagion, Influence

Whether some does something at one time can be predicted from whether their neighbors have already done

- Infectious disease
- Diffusion of innovations
- Diffusion of ideologies

  Pliny (+110): Christianity is a "contagious superstition"

- Not-obviously-infectious conditions (e.g., obesity, loneliness, divorce) ...

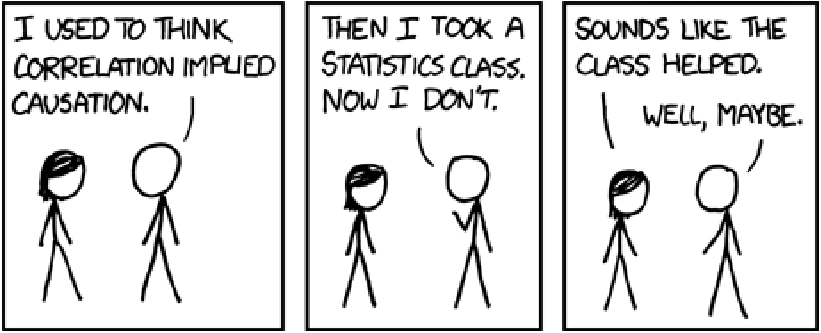
This *can* be due to influence or contagion
Can the same *observational* consequences can follow from latent homophily?

Causal Inference

This is a causal inference question

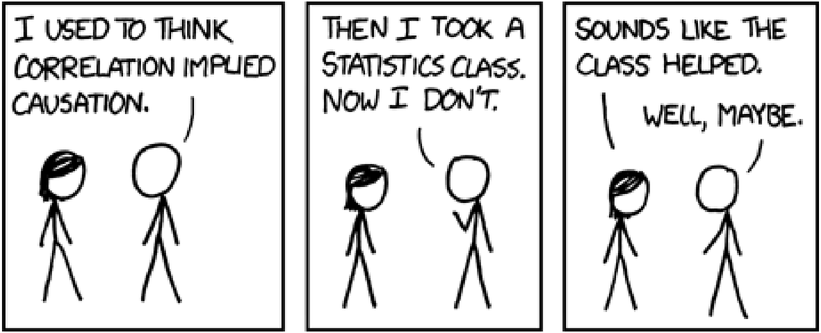
Causal Inference

This is a causal inference question



Causal Inference

This is a causal inference question



"Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'"

Looking Over There

Causal inference becomes a lot clearer once you start drawing graphs (Pearl, 2009; Morgan and Winship, 2007)

Looking Over There

Causal inference becomes a lot clearer once you start drawing graphs (Pearl, 2009; Morgan and Winship, 2007)
nodes = variables, arrows = direct causal influence

Looking Over There

Causal inference becomes a lot clearer once you start drawing graphs (Pearl, 2009; Morgan and Winship, 2007)
nodes = variables, arrows = direct causal influence
Do controls block off indirect paths between variables?

Looking Over There

Causal inference becomes a lot clearer once you start drawing graphs (Pearl, 2009; Morgan and Winship, 2007)
nodes = variables, arrows = direct causal influence
Do controls block off indirect paths between variables?
Do controls *activate* indirect paths?

Looking Over There

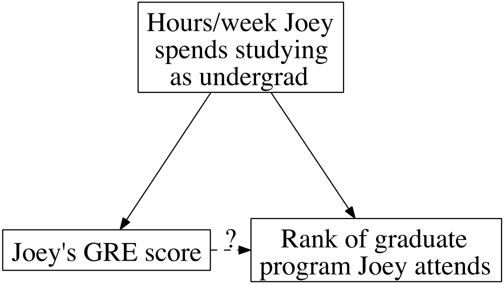
Causal inference becomes a lot clearer once you start drawing graphs (Pearl, 2009; Morgan and Winship, 2007)

nodes = variables, arrows = direct causal influence

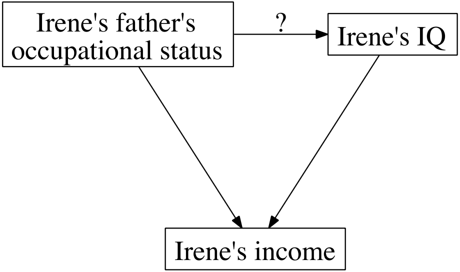
Do controls block off indirect paths between variables?

Do controls *activate* indirect paths?

Separate question: what causal diagrams are compatible with the correlation pattern? (Spirtes *et al.*, 2001)



CONTROL: Conditioning on hours studied lets us estimate the effect of GRE scores on college admission



CONFOUNDING: Conditioning on child's income *makes* child's IQ and father's status dependent

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

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:

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 $Y(i, t)$

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

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)

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:

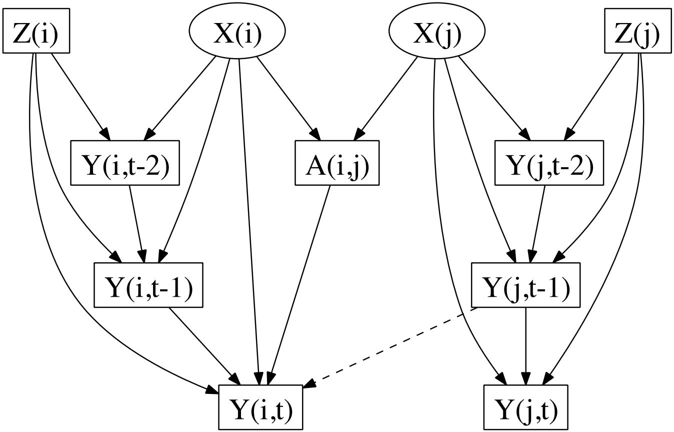
- $Y(i,t-1)$ has a direct influence on $Y(i,t)$
- $X(i)$ has a direct influence on whether/when i adopts
- $Z(i)$ has a direct influence on $Y(i,t)$ (possibly null)
- $Y(j,t-1)$ *may* have a direct influence on $Y(i,t)$, but only if $A(i,j) = 1$

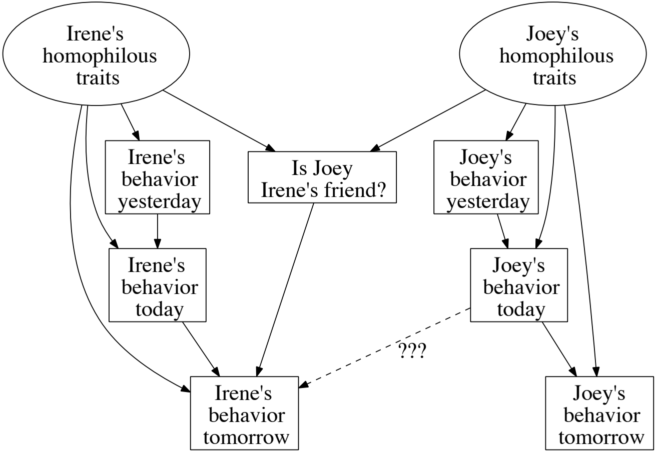
Notation:

- $Y(i, t) =$ 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-2)
A(i,j)
Y(j,t-2)
Y(i,t-1)
Y(j,t-1)
Y(i,t)
Y(j,t)


Irene's homophilous traits
Joey's homophilous traits
Irene's behavior yesterday
Is Joey Irene's friend?
Joey's behavior yesterday
Irene's behavior today
Joey's behavior today
???
Irene's behavior tomorrow
Joey's behavior tomorrow

Contagion Effects are Nonparametrically Unidentifiable

Informally:

1. Joey's behavior yesterday has information about Joey's traits

Contagion Effects are Nonparametrically Unidentifiable

Informally:

1. Joey's behavior yesterday has information about Joey's traits
2. Joey's traits have information about Irene's, *since* they are neighbors

Contagion Effects are Nonparametrically Unidentifiable

Informally:

1. Joey's behavior yesterday has information about Joey's traits
2. Joey's traits have information about Irene's, *since* they are neighbors
3. Irene's traits have information about Irene's behavior today

Contagion Effects are Nonparametrically Unidentifiable

Informally:

1. Joey's behavior yesterday has information about Joey's traits
2. Joey's traits have information about Irene's, *since* they are neighbors
3. Irene's traits have information about Irene's behavior today
4. ∴ Joey's behavior yesterday predicts Irene's behavior today

Contagion Effects are Nonparametrically Unidentifiable

Informally:

1. Joey's behavior yesterday has information about Joey's traits
2. Joey's traits have information about Irene's, *since* they are neighbors
3. Irene's traits have information about Irene's behavior today
4. $\therefore$ Joey's behavior yesterday predicts Irene's behavior today *even if there is no direct causal effect*

Contagion Effects are Nonparametrically Unidentifiable

Informally:

1. Joey's behavior yesterday has information about Joey's traits
2. Joey's traits have information about Irene's, *since* they are neighbors
3. Irene's traits have information about Irene's behavior today
4. $\therefore$ Joey's behavior yesterday predicts Irene's behavior today *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)

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)

The path is not blocked by conditioning on $Y(j,t-2)$

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)

The path is not blocked by conditioning on $Y(j,t-2)$, $Y(i,t-1)$

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)

The path is not blocked by conditioning on $Y(j,t-2)$, $Y(i,t-1)$, $Y(i,t-2)$

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)

The path is not blocked by conditioning on $Y(j,t-2)$, $Y(i,t-1)$, $Y(i,t-2)$ or $Z(i), Z(j)$

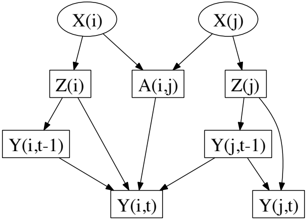
More formally:

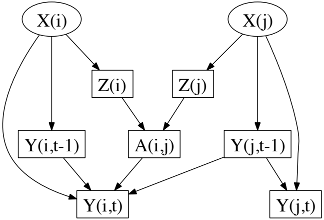
1. $Y(i,t) \leftarrow X(i) \rightarrow A(i,j)$ is a confounding path from $Y(i,t)$ to $A(i,j)$
2. Likewise $Y(j,t-1) \leftarrow X(j) \rightarrow A(i,j)$ is a confounding path from $Y(j,t-1)$ to $A(i,j)$
3. $\therefore$ the direct effect of $Y(j,t-1)$ on $Y(i,t)$ is not identifiable (Pearl, 2009, §3.5,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

The path is not blocked by conditioning on $Y(j,t-2)$, $Y(i,t-1)$, $Y(i,t-2)$ or $Z(i), Z(j)$
Time-varying edges don't help (more spaghetti; cf. Noel and Nyhan (2010))

Getting Identifiability

Parametric assumptions *might* 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, $i \rightarrow j$, $j \nrightarrow i$

The Argument from Asymmetry

Focus on unreciprocated edges, $i \rightarrow j$, $j \not\rightarrow i$

IRENE: *Joey is my friend!*
JOEY: *Irene who?*

The Argument from Asymmetry

Focus on unreciprocated edges, $i \rightarrow j$, $j \not\rightarrow i$

> IRENE: *Joey is my friend!*
> JOEY: *Irene who?*

Suppose $Y(i,t)|Y(j,t-1)) \not\sim Y(j,t)|Y(i,t-1)$
Doesn't this argue for direct influence?

The Argument from Asymmetry

Focus on unreciprocated edges, $i \rightarrow j$, $j \not\rightarrow i$

> IRENE: *Joey is my friend!*
> JOEY: *Irene who?*

Suppose $Y(i,t)|Y(j,t-1)) \not\sim Y(j,t)|Y(i,t-1)$
Doesn't this argue for direct influence?
Sounds plausible…

The Argument from Asymmetry

Focus on unreciprocated edges, $i \rightarrow j$, $j \not\rightarrow i$

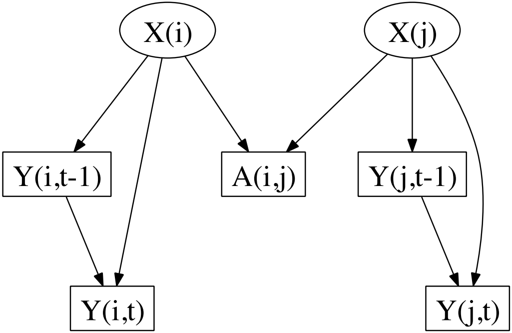
> IRENE: *Joey is my friend!*
> JOEY: *Irene who?*

Suppose $Y(i,t)|Y(j,t-1)) \not\sim Y(j,t)|Y(i,t-1)$
Doesn't this argue for direct influence?
Sounds plausible...
...fails if senders and receivers have systematically different values of X, with different local relations to Y

Toy Example

Try to predict $Y(i,t)$ from $Y(j,t)$ and vice versa when
$A_{ij} = 1, A_{ji} = 0$
$X(i) \sim \mathcal{U}(0,1)$
Edges form with probability $\propto \mathrm{logit}^{-1}(-3|X(i) - X(j)|)$
i nominates j from among neighbors, $\propto \mathrm{logit}^{-1}(-|X(j) - 0.5|)$

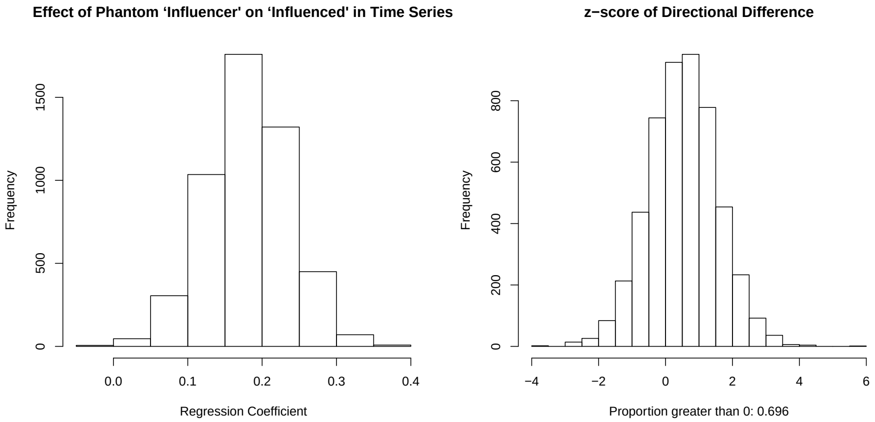
$$
\begin{aligned}
Y(i,0) &= (X(i) - 0.5)^3 + \mathcal{N}(0,(0.02)^2) \\
Y(i,1) &= Y(i,0) + 0.3X_i + \mathcal{N}(0,(0.02)^2)
\end{aligned}
$$



Causal graph of the model with no contagion, but asymmetry in regression coefficients

Results:

- $Y(i, 1)$ is well-predicted from $Y(j, 0)$
- *Nominees* are disproportionately in the middle; $i \rightarrow j, j \not\rightarrow i$ suggests i is more peripheral
- For asymmetric pairs, regression of sender on receiver differs from that of receiver on sender


Effect of Phantom 'Influencer' on 'Influenced' in Time Series
Frequency
0
500
1000
1500
0.0
0.1
0.2
0.3
0.4
Regression Coefficient
z−score of Directional Difference
Frequency
0
200
400
600
800
−4
−2
0
2
4
6
Proportion greater than 0: 0.696

Making homophily and contagion look like causation

A central theme of social science:
Long-term, hard-to-change social/economic status variables explain short-term, malleable cultural / political / consumer variables
Culture and choices express (reflect, serve, ...) social/economic interests or experiences

Making homophily and contagion look like causation

A central theme of social science:
Long-term, hard-to-change social/economic status variables explain short-term, malleable cultural / political / consumer variables
Culture and choices express (reflect, serve, ...) social/economic interests or experiences

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

Quantitatively: use differences in demographics to predict differences in wedding gowns (or survey answers)

What's the evidence?

What's the evidence?

- The stories sound good

What's the evidence?

- The stories sound good
- Casual empiricism

What’s the evidence?

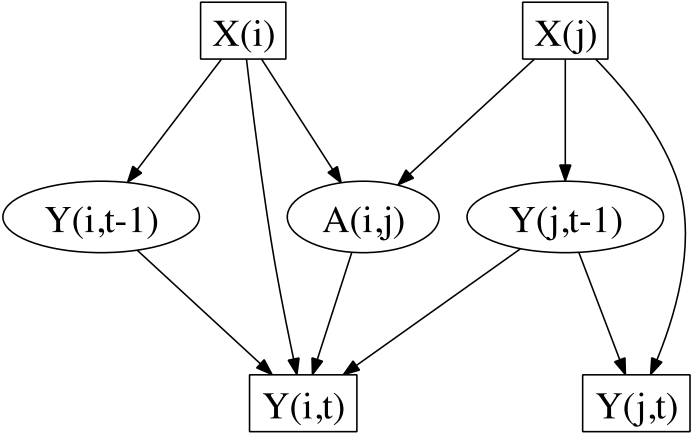
- The stories sound good
- Casual empiricism
- Correlation/regression analyses; cultural choices are predictable from social positions (e.g. Bourdieu (1984))

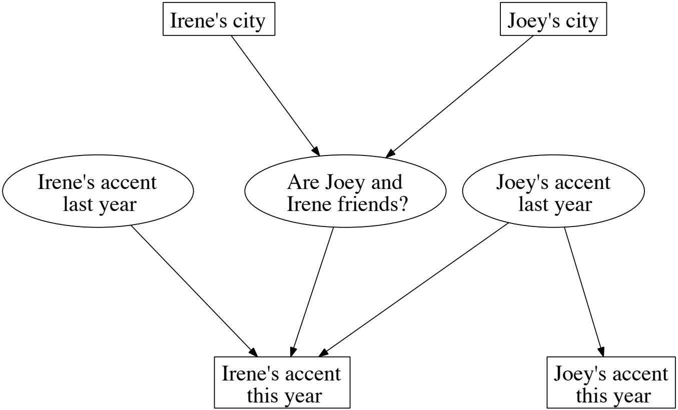
Probably true a lot of the time

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 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)


Irene's city
Joey's city
Irene's accent last year
Are Joey and Irene friends?
Joey's accent last year
Irene's accent this year
Joey's accent this year

More Confounding

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

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. $A(i,j)$

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, $Y(j, t-1)$

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 $A(i,j)=1$

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**

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

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

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")

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)

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")

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")
- $Y(\cdot, 0)$ = Bernoulli(1/2) process

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")
- $Y(\cdot,0) =$ Bernoulli(1/2) process
    1. At each t, pick a random i, and a random neighbor j

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")
- $Y(\cdot,0)$ = Bernoulli(1/2) process
    1. At each t, pick a random i, and a random neighbor j
    2. Set $Y(i,t) = Y(j,t-1)$

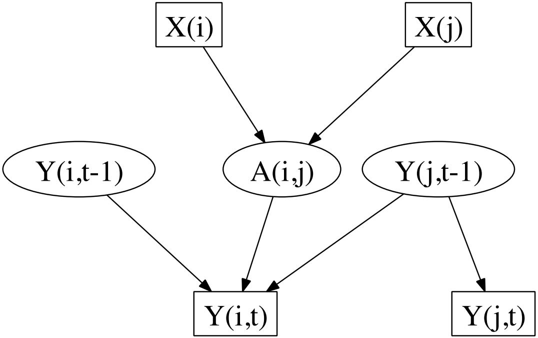
Caricature Neutral Model of Cultural Evolution

- $X(i) =$ 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")
- $Y(\cdot, 0) =$ Bernoulli(1/2) process
    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)

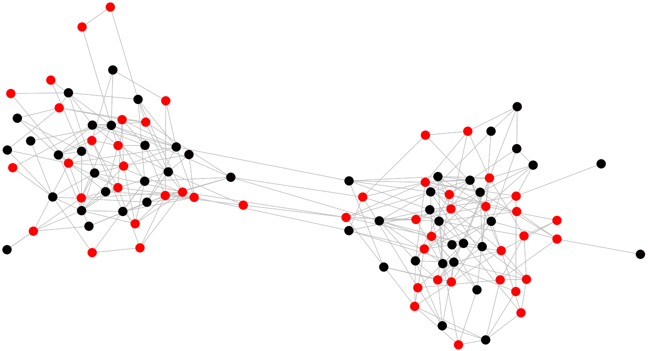
Caricature Neutral Model of Cultural Evolution

- $X(i)$ = 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")
- $Y(\cdot, 0)$ = Bernoulli(1/2) process
    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)

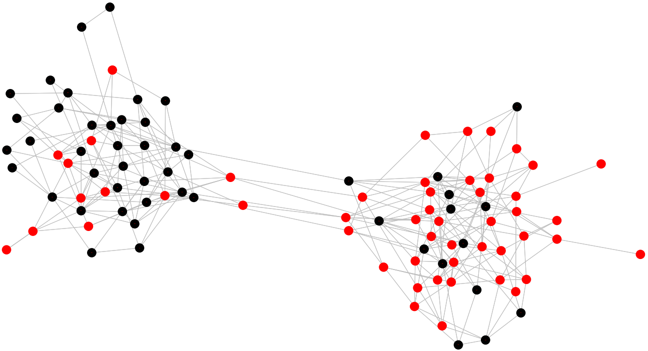
(= "voter model" of statistical mechanics)



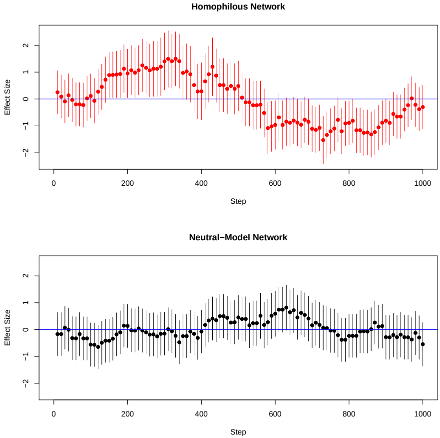
Graph for the voter model of neutral cultural evolution



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



After 300 updates



Coefficients for logistic regression of choice on status, $\pm$ 95% confidence intervals. Red, homophilous network; black, matched non-assortative network

Contagion + Homophily Looks Like Causation

- Neutral diffusion + homophily looks like a real connection between social status and cultural choices

Contagion + Homophily Looks Like Causation

- Neutral diffusion + homophily looks like a real connection between social status and cultural choices
- Problem is *not* the ecological or “red-state/blue-state” fallacy (not using aggregated data)

Contagion + Homophily Looks Like Causation

- Neutral diffusion + homophily looks like a real connection between social status and cultural choices
- Problem is *not* the ecological or “red-state/blue-state” fallacy (not using aggregated data)
- Problem is *not* using the complete population instead of a random sample

Contagion + Homophily Looks Like Causation

- Neutral diffusion + homophily looks like a real connection between social status and cultural choices
- Problem is *not* the ecological or “red-state/blue-state” fallacy (not using aggregated data)
- Problem is *not* using the complete population instead of a random sample
- Problem *is* that choices are not independent conditional on statuses

Contagion + Homophily Looks Like Causation

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

What To Do?

How can we go forward with studying contagion when there is homophily?

- Experiment: on Y or A or both

What To Do?

How can we go forward with studying contagion when there is homophily?

- Experiment: on Y or A or both
  this is Science, and is Hard

What To Do?

How can we go forward with studying contagion when there is homophily?

- Experiment: on Y or A or both
  this is Science, and is Hard
- Homophily on X is no problem if we condition on X

What To Do?

How can we go forward with studying contagion when there is homophily?

- Experiment: on Y or A or both
  this is Science, and is Hard
- Homophily on X is no problem if we condition on X
  $\therefore$ figure out what X is and measure it

What To Do?

How can we go forward with studying contagion when there is homophily?

- Experiment: on Y or A or both
  this is Science, and is Hard
- Homophily on X is no problem if we condition on X
  $\therefore$ figure out what X is and measure it
  this is Science, and is Hard

What To Do?

How can we go forward with studying contagion when there is homophily?

- Experiment: on Y or A or both
  this is Science, and is Hard
- Homophily on X is no problem if we condition on X
  $\therefore$ figure out what X is and measure it
  this is Science, and is Hard
- Bounds: even if we can't point-identify, maybe we can pin down a range

What To Do?

How can we go forward with studying contagion when there is homophily?

- Experiment: on Y or A or both
  this is Science, and is Hard
- Homophily on X is no problem if we condition on X
  $\therefore$ figure out what X is and measure it
  this is Science, and is Hard
- Bounds: even if we can't point-identify, maybe we can pin down a range
- Clustering: figure out X from the social network

Bounds and Partial Identification

Unidentifiable parameter $\equiv$ multiple values of the parameter yield the *same* observational distribution

Bounds and Partial Identification

Unidentifiable parameter $\equiv$ multiple values of the parameter yield the *same* observational distribution
$\therefore$ even infinite data does not pin down the parameter

Bounds and Partial Identification

Unidentifiable parameter $\equiv$ multiple values of the parameter yield the *same* observational distribution

$\therefore$ even infinite data does not pin down the parameter

Partial identification (Manski, 2007): range of parameter values yielding one distribution might be limited

Bounds and Partial Identification

Unidentifiable parameter $\equiv$ multiple values of the parameter yield the *same* observational distribution
$\therefore$ even infinite data does not pin down the parameter
Partial identification (Manski, 2007): range of parameter values yielding one distribution might be limited
$\therefore$ infinite data *bounds* the parameter

Bounds and Partial Identification

Unidentifiable parameter $\equiv$ multiple values of the parameter yield the *same* observational distribution
$\therefore$ even infinite data does not pin down the parameter
Partial identification (Manski, 2007): range of parameter values yielding one distribution might be limited
$\therefore$ infinite data *bounds* the parameter
Could we bound contagion effects when there is latent homophily?

Bounds and Partial Identification

Unidentifiable parameter $\equiv$ multiple values of the parameter yield the *same* observational distribution

$\therefore$ even infinite data does not pin down the parameter

Partial identification (Manski, 2007): range of parameter values yielding one distribution might be limited

$\therefore$ infinite data *bounds* the parameter

Could we bound contagion effects when there is latent homophily?

Work in progress, only negative results so far

Actual contagion effect cannot be bounded from regression coefficient even in linear Gaussian model

Partial Control by Clustering?

Can we make the latent trait manifest?

Partial Control by Clustering?

Can we make the latent trait manifest?
Latent homophily $\Rightarrow$ you tend to resemble your neighbors

Partial Control by Clustering?

Can we make the latent trait manifest?
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.

Partial Control by Clustering?

Can we make the latent trait manifest?
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

Partial Control by Clustering?

Can we make the latent trait manifest?
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
Try using community membership as a proxy for X

Partial Control by Clustering?

Can we make the latent trait manifest?
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
Try using community membership as a proxy for X
Removes confounding if estimated communities are a sufficient statistic for X

Partial Control by Clustering?

Can we make the latent trait manifest?
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
Try using community membership as a proxy for X
Removes confounding if estimated communities are a sufficient statistic for X

When does that hold?

What about other block models?

Tighter bounds, even if not identified?

Conclusion

1. Social linkage creates causal confounding
2. Homophily + causal influence looks like contagion
3. Homophily + contagion looks like causal influence
4. *May* be possible to *limit* confounding

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

Christakis, Nicholas A. and James H. Fowler (2007). "The Spread of Obesity in a Large Social Network over 32 Years." *The New England Journal of Medicine*, **357**: 370–379. URL http://content.nejm.org/cgi/content/abstract/357/4/370.

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. URL http://www.jstor.org/pss/2298123.

— (2007). *Identification for Prediction and Decision*. Cambridge, Massachusetts: Harvard University Press.

Morgan, Stephen L. and Christopher Winship (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge, England: Cambridge University Press.

Noel, Hans and Brendan Nyhan (2010). "The "Unfriending" Problem The Consequences of Homophily in Friendship Retention for Causal Estimates of Social Influence." Online preprint. URL http://www-personal.umich.edu/~bnyhan/unfriending.pdf.

Pearl, Judea (2009). *Causality: Models, Reasoning, and*

Inference. Cambridge, England: Cambridge University Press, 2nd edn.

Shalizi, Cosma Rohilla and Andrew C. Thomas (2011). "Homophily and Contagion Are Generically Confounded in Observational Social Network Studies." *Sociological Methods and Research*, **40**: 211–239. URL http://arxiv.org/abs/1004.4704. doi:10.1177/0049124111404820.

Sperber, Dan (1996). *Explaining Culture: A Naturalistic Approach*. Oxford: Basil Blackwell.

Spirtes, Peter, Clark Glymour and Richard Scheines (2001). *Causation, Prediction, and Search*. Cambridge, Massachusetts: MIT Press, 2nd edn.

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`.