SOSC 4300/5500: Homophily, Network, and Causal Inference

Han Zhang

Nov 24, 2020

Outline

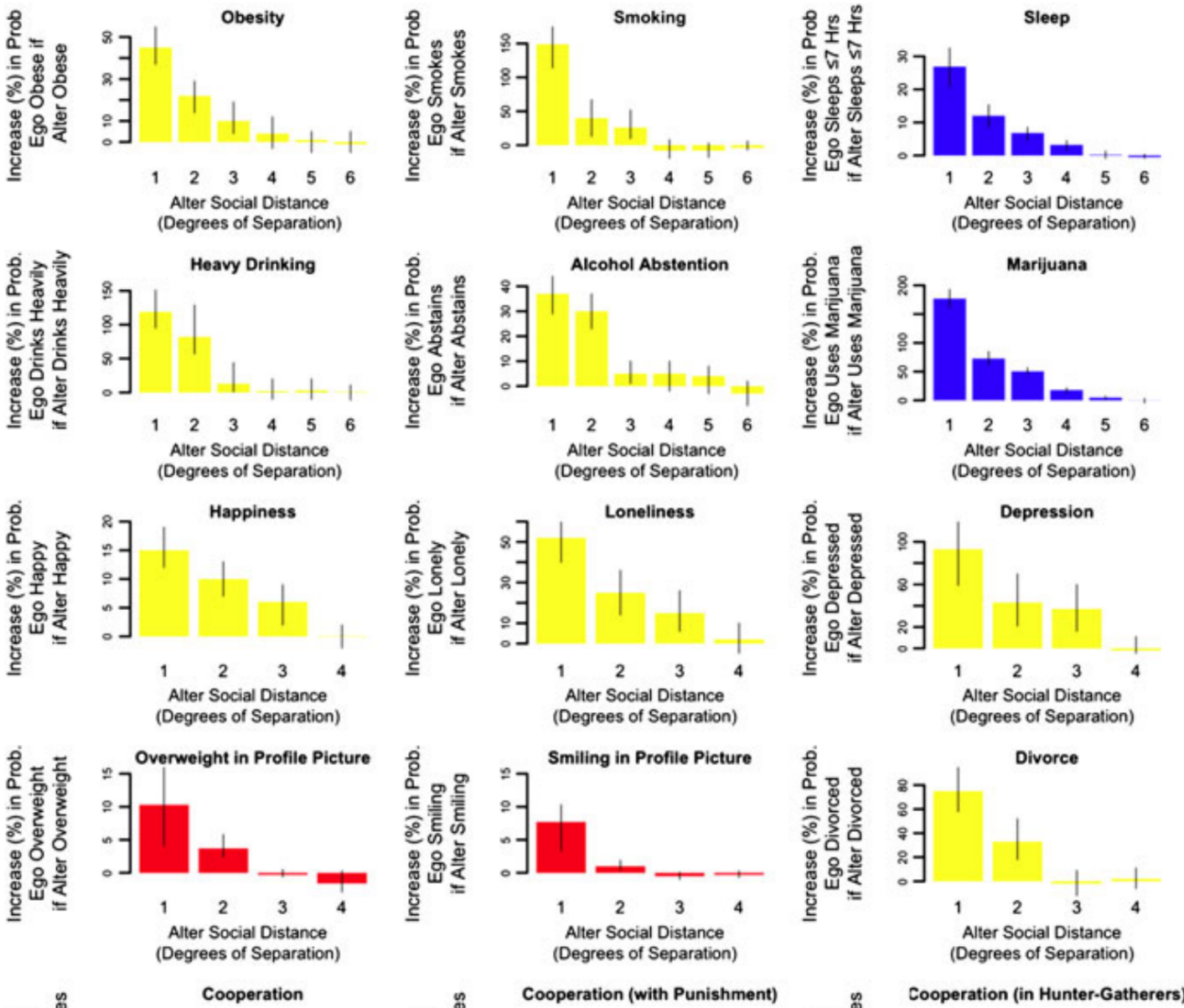
Empirical work on diffusion

- Last week we talked about mathematical models of diffusion
  - Simple contagion on small-world network
  - Complex contagion on complex networks
- Let us move on to empirical results regarding diffusion
- First, focus on observational data

Friends are similar to each other

- Countless evidence suggests that friends are usually more similar to each other than to random others
  - In mathematical terms, the correlation between friends' attitutes tend to be higher than the correlation between two random nodes
- List to similar music; watch similar sports
- Similar demographics (age, education, etc)
- others?

Friends are similar to each other



Diffusion of Obesity: social influence

- Nicholas A. Christakis and James H. Fowler, *The Spread of Obesity in a Large Social Network over 32 Years*, New England Journal of Medicine **357** (2007), no. 4, 370–379
- `https://www.youtube.com/watch?v=8aEtyRD1j5U`
- Christakis and Fowler think it's because of social influence:
  - you follow your friends' behaviors
  - so obesity diffuses along social networks
- Other explanations?

Alternative mechanism 1: homophily

- Homophily: people are more likely to befriend with others similar to themselves
  - Primary homophily: overweight people become friends with other overweight people
  - Second homophily: overweight people share some other characteristics (e.g., lower-income family in the US), which drives homophily
    - note: this one can be observed or unobserved
- This is a type of selection bias
- Miller McPherson, Lynn Smith-Lovin, and James M. Cook, *Birds of a Feather: Homophily in Social Networks*, Annual Review of Sociology **27** (2001), 415–444

Contagion vs. Homophily

- selection versus influence, or contagion vs. homophily
- And it's a general problem, beyond the obesity setting
- Examples:
  - Friends are similar in their music taste
  - Friends usually consume similar political news (so-called filter bubbles or echo chambers)
  - Smoking/drinking
  - Emotion
  - Political party
  - Divorce
  - Else?

Alternative mechanism 2: contextual effect

- They just happy to share the same environment
  - A type of omitted variable bias, or confounding biases, or common external causes
- E.g., live in poor neighborhood with only fast food restaurants
- Contextual bias is *relatively* easier to control; get more measures on confounding variables

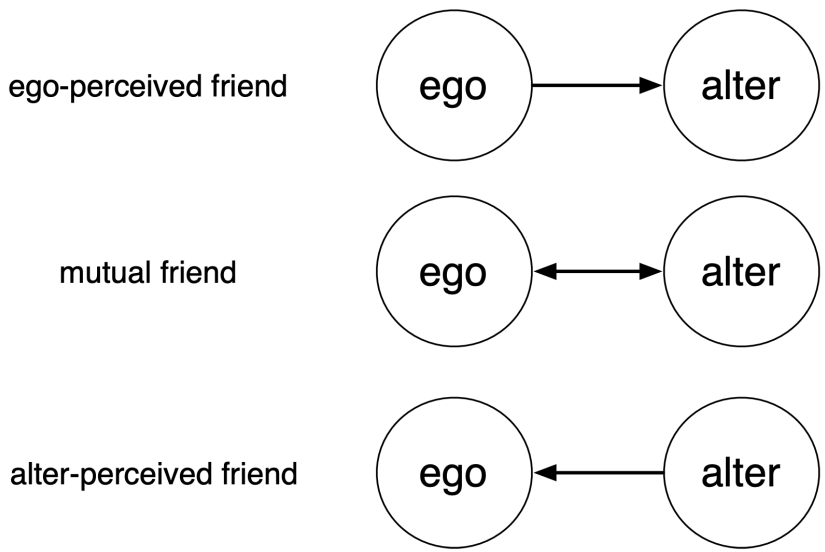
Christakis and Fowler's statistical model

- Use statistical modeling
- Regress whether ego is obese at time $t+1$ based on:
- alter obese at t
  - outcome of interest
- ego obese at t
  - Control for autocorrelation
- alter obese at $t+1$
  - Control for homophily
- ego demographic variables (age, gender, education, etc.)
  - Control for observed confoundings

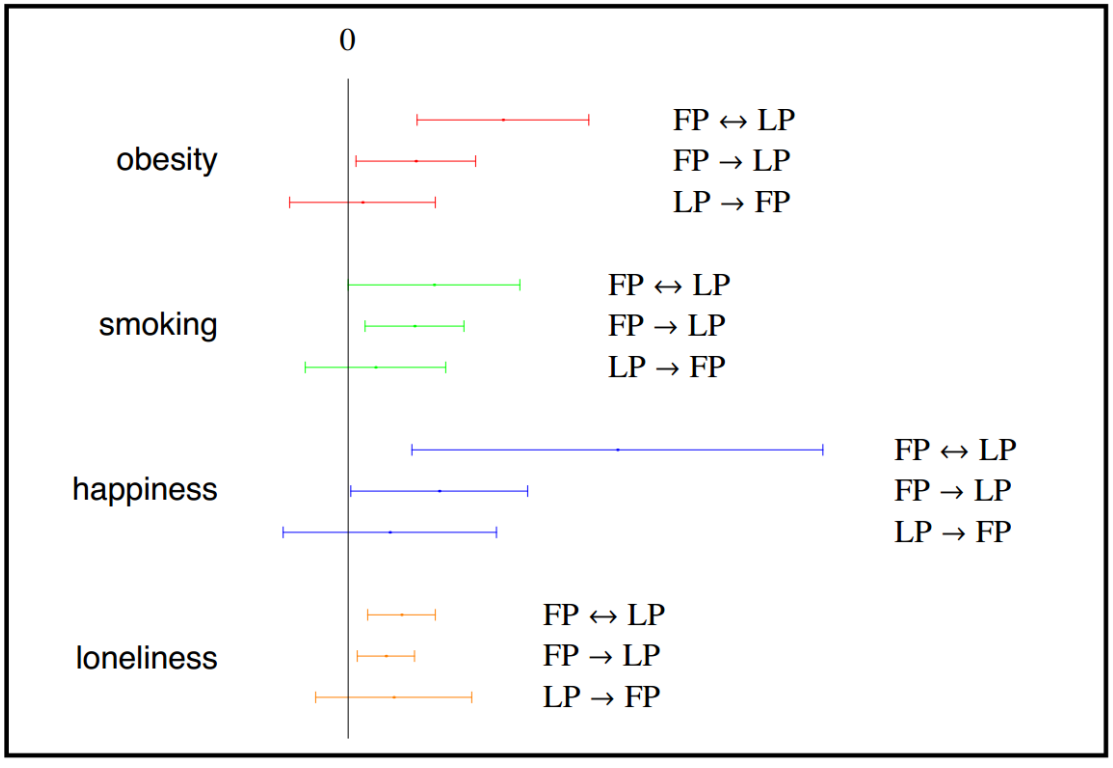
$$Y_{t+1}^{\text{ego}} = \alpha + \beta_1 y_t^{\text{ego}} + \beta_2 y_{t+1}^{\text{alter}} + \beta_3 y_t^{\text{alter}} + \sum_{i=1}^{k} \gamma_i x_i \qquad (1)$$

Unobserved factors

- Even after controlling for many things, we may still have unobserved factors
- To rule out unobserved shared environments: using directionality of ties:
  - ego and alter share similar unobserved factors (similar to twin studies)



Unobserved factor case



- FP is ego and LP is alter
- Differences between the three is often not statistically different

"Towards Responsible Just-So Story Telling"

- Cosma Rohilla Shalizi and Andrew C. Thomas, *Homophily and Contagion Are Generically Confounded in Observational Social Network Studies*, arXiv:1004.4704 (2010)

  Contagion effects are nonparametrically unidentifiable in the presence of unobserved secondary homophily

- Generally, it is challenging to estimate causal effect from purely observational data
- Christakis and Fowler's data are already better than many others have (time-series and directionarlity)

An area of debates

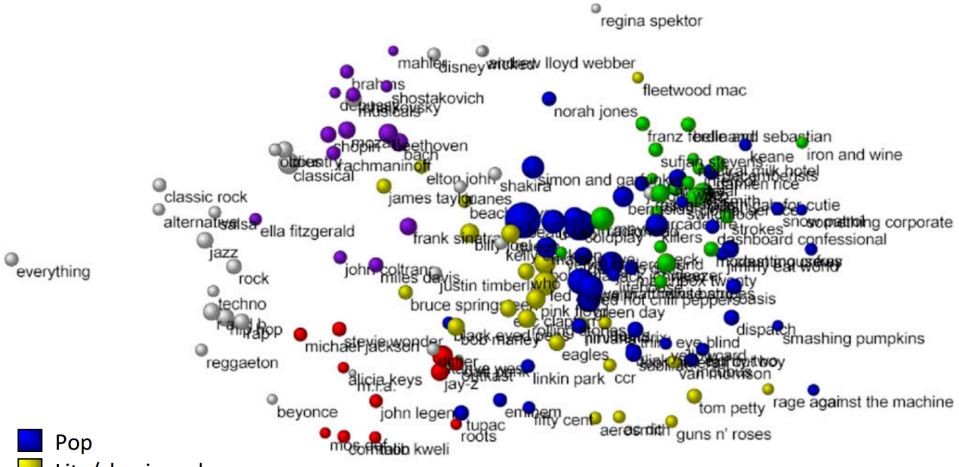
Tyler J. VanderWeele, Elizabeth L. Ogburn, and Tchetgen Eric J. Tchetgen, *Why and When "Flawed" Social Network Analyses Still Yield Valid Tests of no Contagion*, Statistics, Politics, and Policy **3** (2012), no. 1

> *Social network analyses of the type employed by Christakis and Fowler will still yield valid tests of the null of no social contagion, even though estimates and confidence intervals may not be valid.*

Diffusion of Music Taste, or no?

- Kevin Lewis, Marco Gonzalez, and Jason Kaufman, *Social selection and peer influence in an online social network*, Proceedings of the National Academy of Sciences **109** (2012), no. 1, 68–72

  Our data are based on the Facebook activity of a cohort of students at a diverse US college (n = 1,640 at wave 1). Beginning in March 2006 (the students' freshman year) and repeated annually through March 2009 (the students senior year)



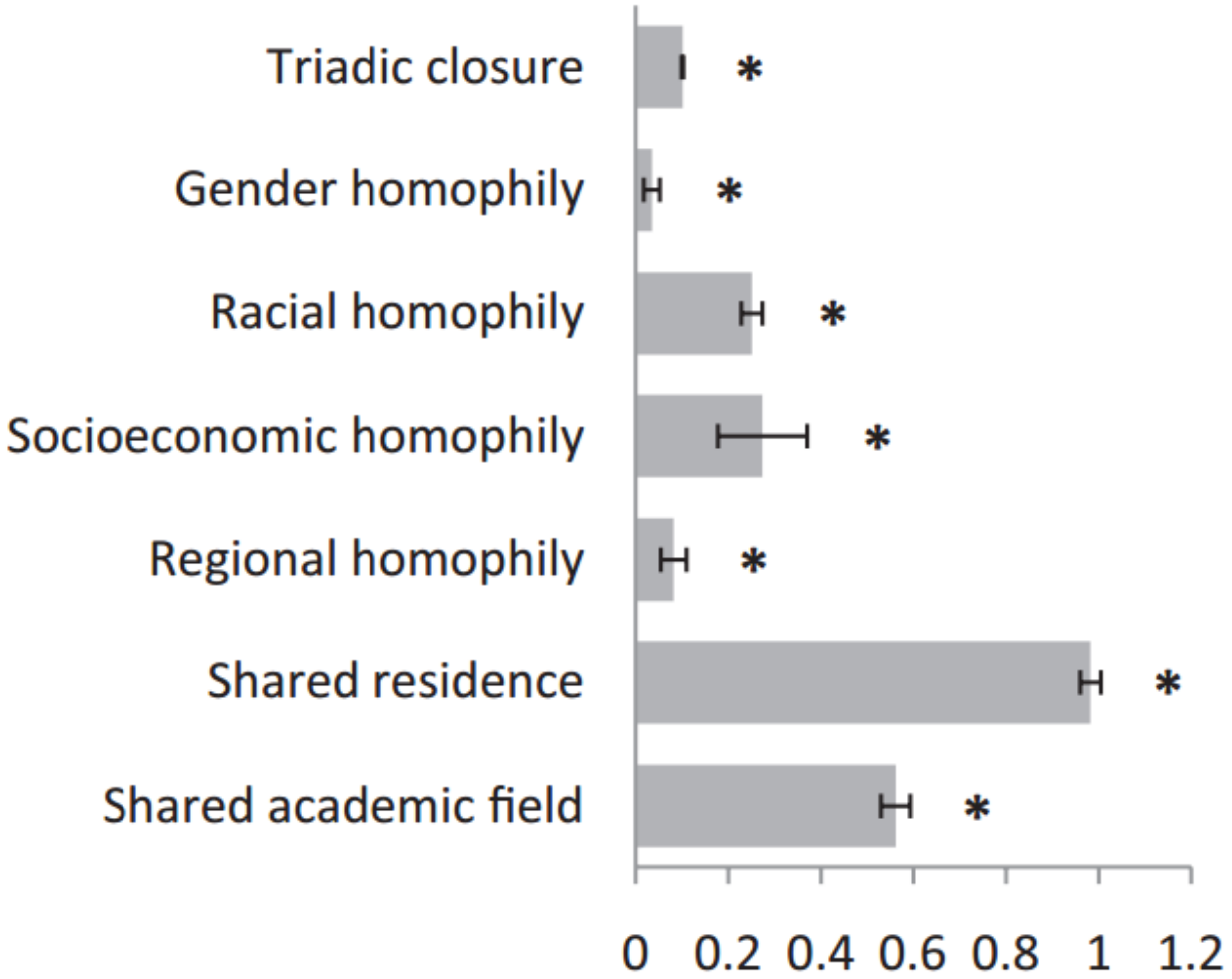
Social influence vs. homophily again

- Are similar music taste between friends
  - social influence
  - homophily?
- A differnet approach: stochastic actor-oriented network models (SAOM)
  - Krzysztof Nowicki and Tom A. B. Snijders, *Estimation and Prediction for Stochastic Blockstructures*, Journal of the American Statistical Association **96** (2001), no. 455, 1077–1087
  - `https://www.stats.ox.ac.uk/~snijders/siena/`

SOAM

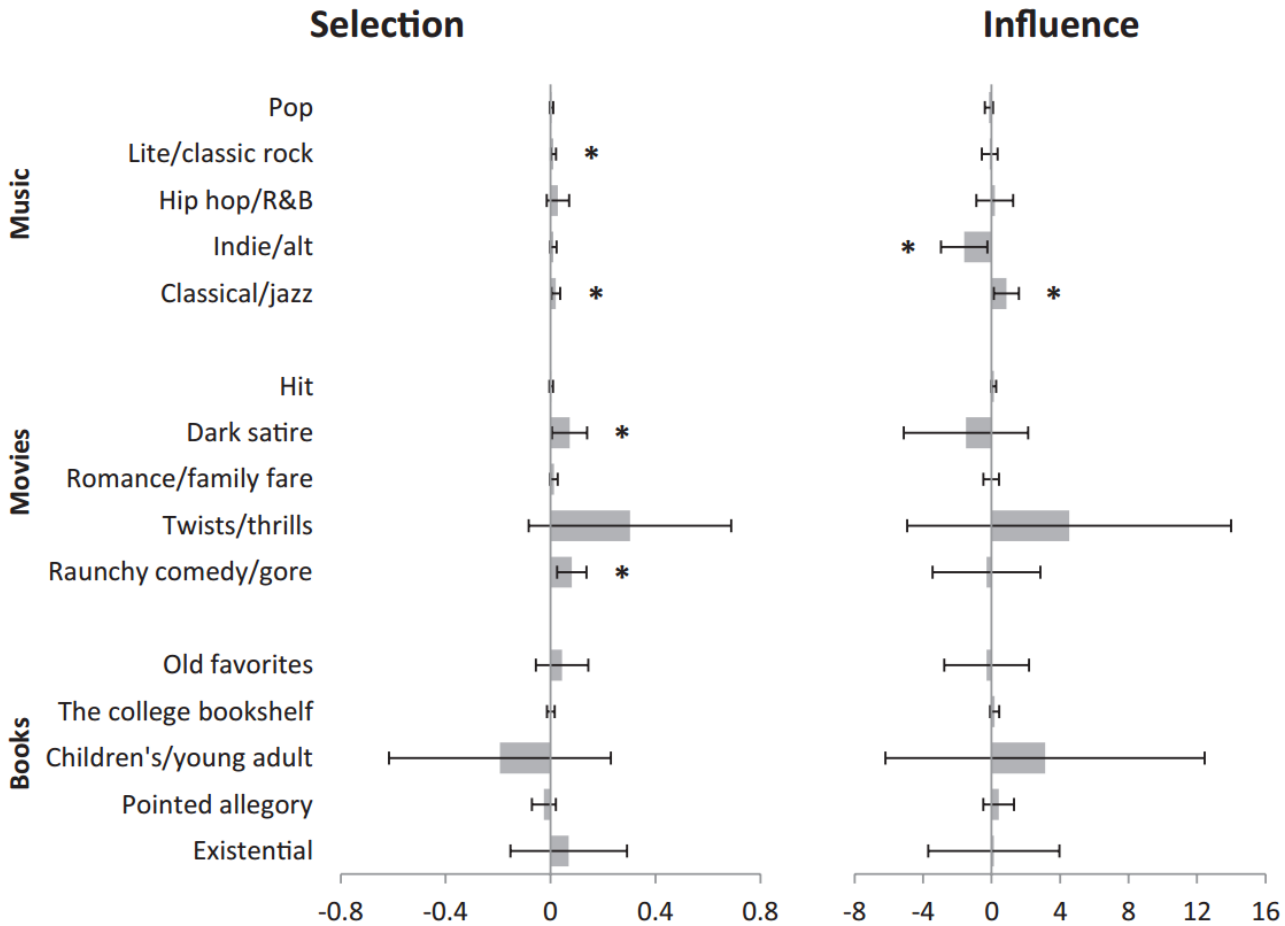
- Key intuitions: say if we want to test whether there is contagion effect
- Simulate what network would looks like, if there were no social influence (other things equal)
- Then compared observed network with simulated network to get the net contagion effect
- This is quite different from the linear regression approach, where observations are assumed to be independently distributed.
- `https://movie-usa.glencoesoftware.com/video/10.1073/pnas.1811388115/video-1`

Results



- Shared space is the mostly salient predictor

Results



- Some evidence for selection/homophily
- But no evidence for social influence

ERGM

- There is a separate branch of statistical analysis that simulates networks and then estimate effects
- Exponential random graph models (ERGM)
- Per Block, Johan Koskinen, James Hollway, Christian Steglich, and Christoph Stadtfeld, *Change we can believe in: Comparing longitudinal network models on consistency, interpretability and predictive power*, Social Networks **52** (2018), 180–191
- SOAM slightly better, but "both models perform poorly in out-of-sample prediction compared to trivial predictive models."

ERGM's problems

- Michael Schweinberger and Mark S. Handcock, *Local dependence in random graph models: Characterization, properties and statistical inference*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **77** (2015), no. 3, 647–676
- Bad predictions

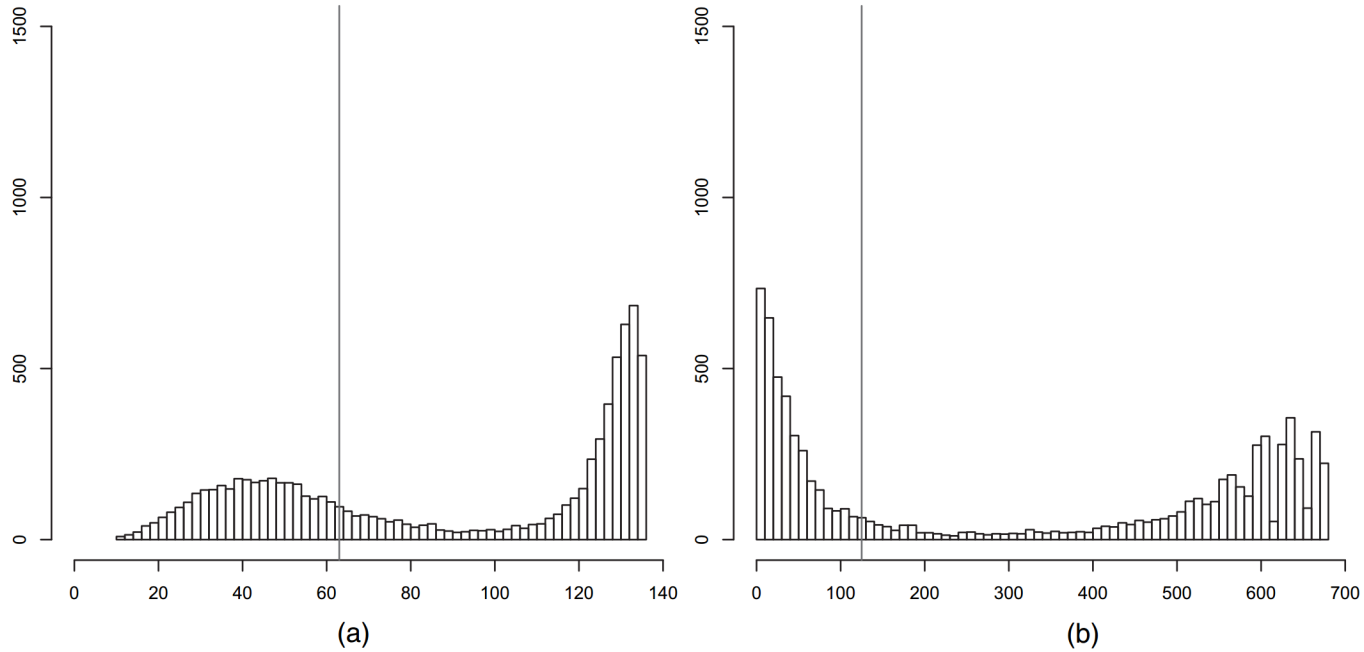
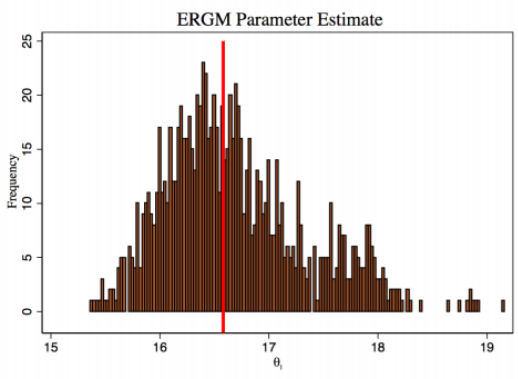


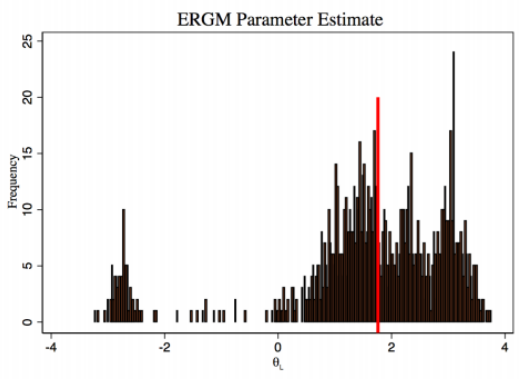
Fig. 7. Terrorist network: posterior predictions of the number of (a) edges and (b) triangles under the global triangle model (|, observed numbers): although the number of edge variables $N = 136$ is not large, the

ERGM's problems

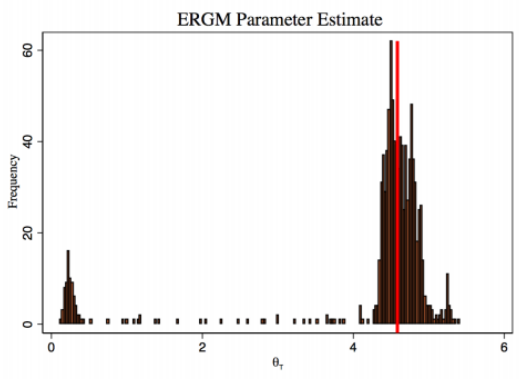
- The poor predictions lead to inconsistent estimator: estimate does not converge to true value when we have more and more observations



(A) Isolate Parameter Estimates



(B) Link Parameter Estimates



(C) Triangle Parameter Estimates

Take-away messages

- Friends are similar to each other
- But it may be due to different reasons
  - homophily
  - contagion/social influence
  - common environment
- Contagion effect is difficult to estimate in observational data when people can choose their friends and may be exposed to environmental changes that we don't measure
- To do it right, you need to take more statistics class
  - and even the best models so far have problems
- Or alternatively, consider running experiments
- Both are frontiers of science, now!

Randomized controlled experiments

- Randomized controlled experiments
  - Split subjects into treated/control groups randomly
  - Provide treatment to treatd users, not to control users
  - Compare difference in means
- Simple and straightforward, compared with observational studies

Two types of experiments

- Lab experiment
  - Pros: full control
  - Cons: lack of external validity; full of convenient samples
- Field experiment
  - Pros: external validity;
  - Cons: take a lot of resources
- Internet era has led to online lab/field experiment

Emotion Contagion

- Hypothesis: seeing lots of happy things can make you happy.
- Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, *Experimental evidence of massive-scale emotional contagion through social networks*, Proceedings of the National Academy of Sciences **111** (2014), no. 24, 8788–8790
- Experimental design (online field experiment)
  - 700,000 people
  - Three groups:
    - positive posts in Facebook News Feed reduced (by 10% to 90%)
    - negative posts in Facebook News Feed reduced
    - control
  - Post scored as positive or negative based on LIWC dictionary
  - Outcome: proportion of words posted that were positive or negative in 1 week of experiment

Emotion Contagion

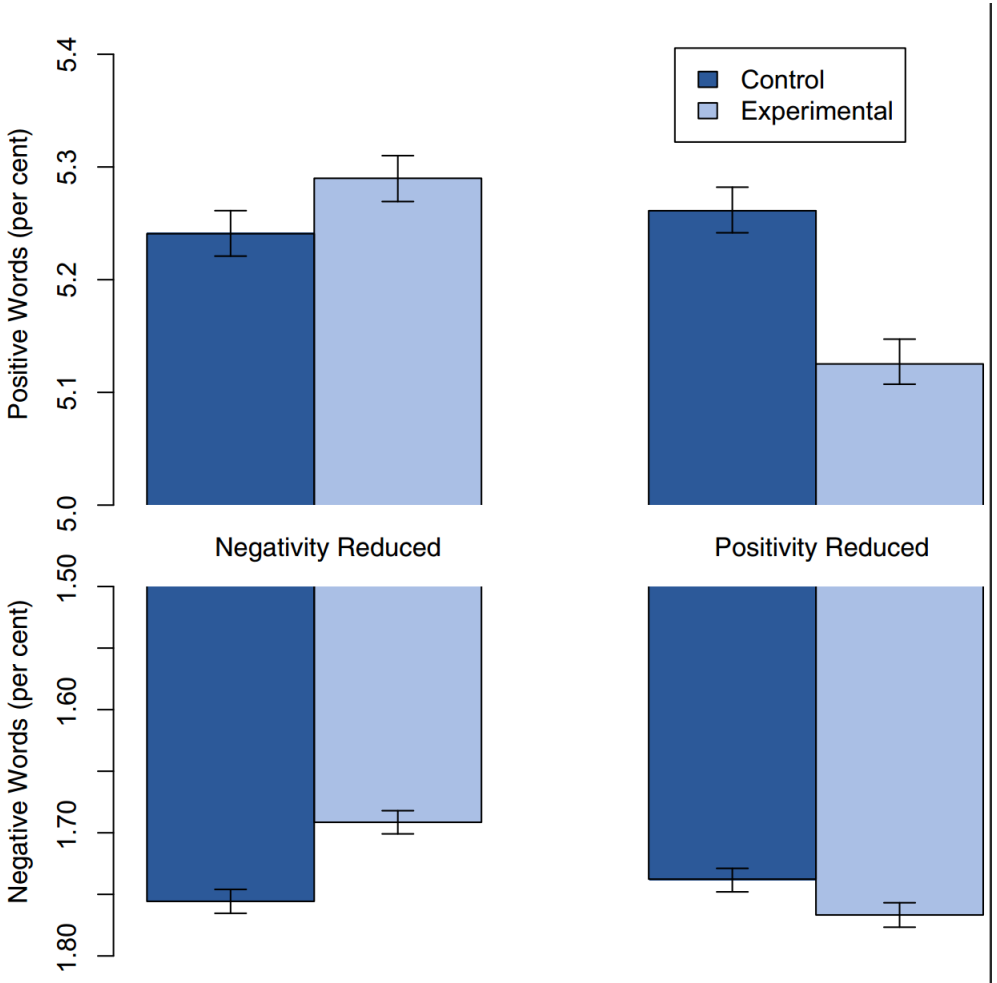


Fig. 1. Mean number of positive (*Upper*) and negative (*Lower*) emotion words (percent) generated people, by condition. Bars represent standard errors.

Questions

- Is posting on Facebook a good measure of how we feel?
- Is LIWC dictionary methods a good way to quantify emotional content of a Facebook post?

Ethical concerns

- Most importantly, should this experiment ever have happened?
- Facebook reveals news feed experiment to control emotions;Protests over secret study involving 689,000 users in which friends' postings were moved to influence moods
- Stop complaining about the Facebook study. It's a golden age for research
- A collection of articles if you are further interested

`http://laboratorium.net/archive/2014/06/30/the_facebook_emotional_manipulation_study_source`

Simple and Complex contagion

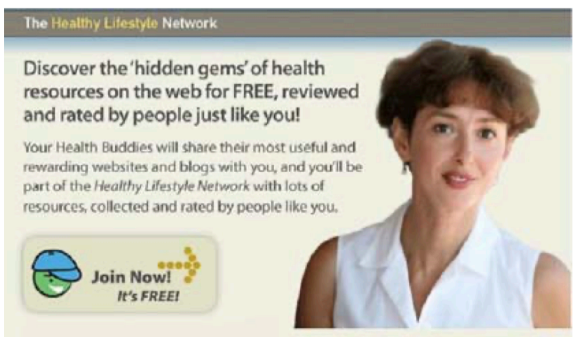
- Last week, we learn the differences bewteen simple and complex contagion

	Regular network (lattice)	Small-world network
Simple contagion	slow	fast
Complex contagion	fast	slow

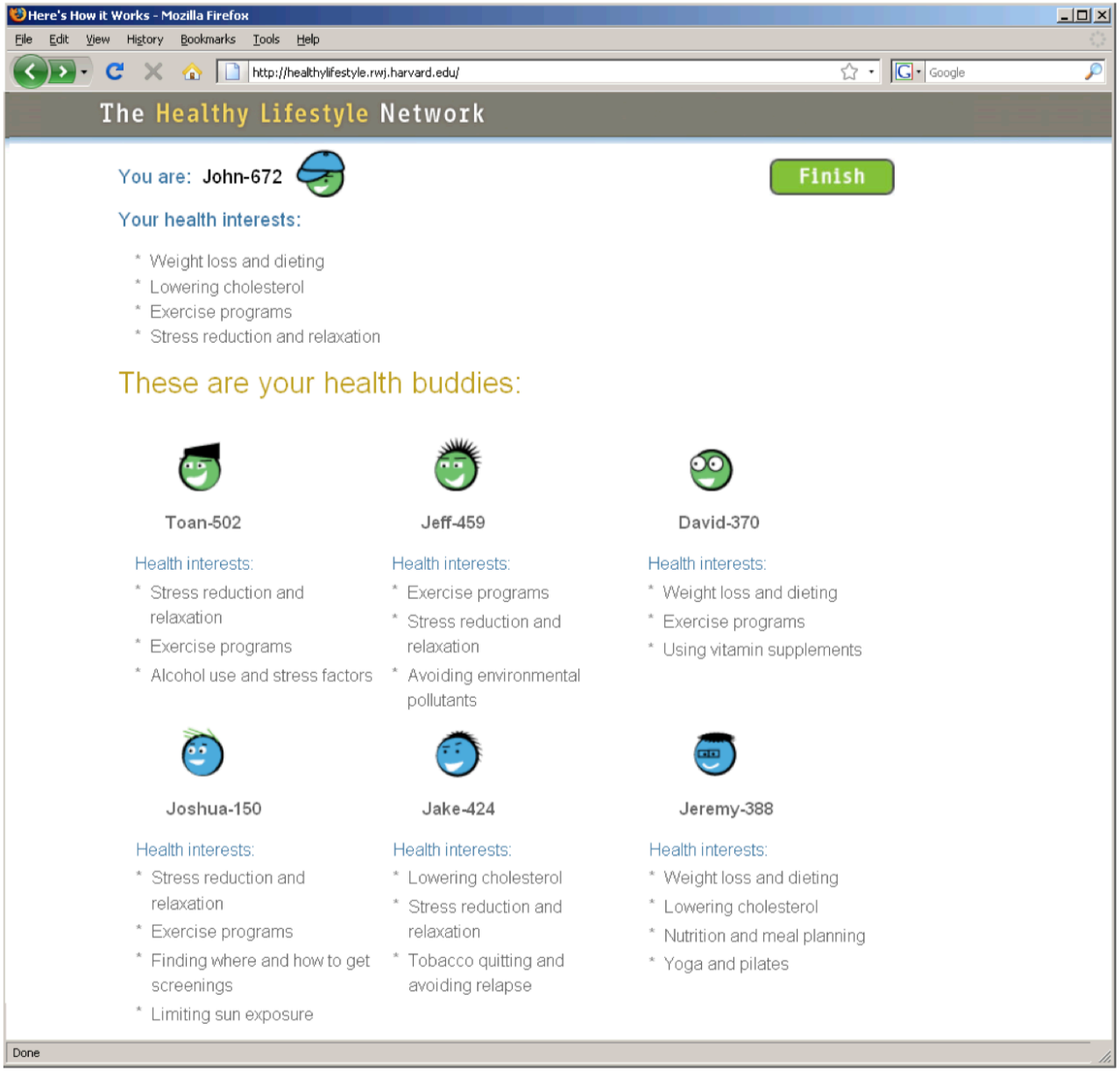
- Can this theory be empirically tested?
- Damon Centola, *The Spread of Behavior in an Online Social Network Experiment*, Science **329** (2010), no. 5996, 1194–1197
- This is an example of online lab experiment

Experiment recruitment

- "I created an Internet-based health community, containing 1528 participants recruited from health-interest World Wide Web sites (13)."



What this health community looks like?



Experiment design

- A random node is reserved for researchers (the seed nodes)
- Respondents were assigned to other nodes and randomly to two conditions

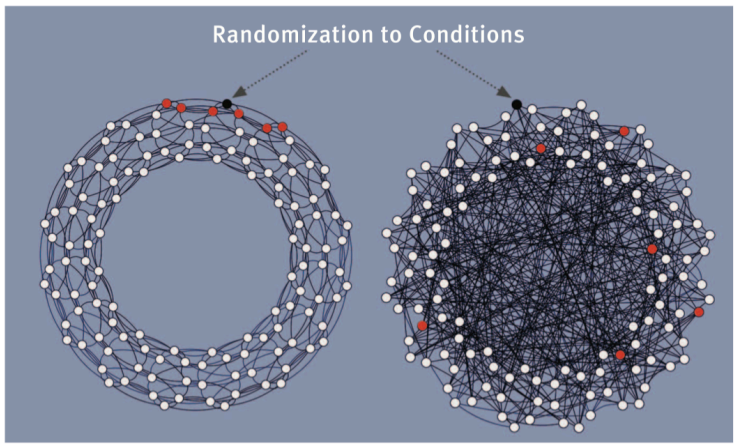
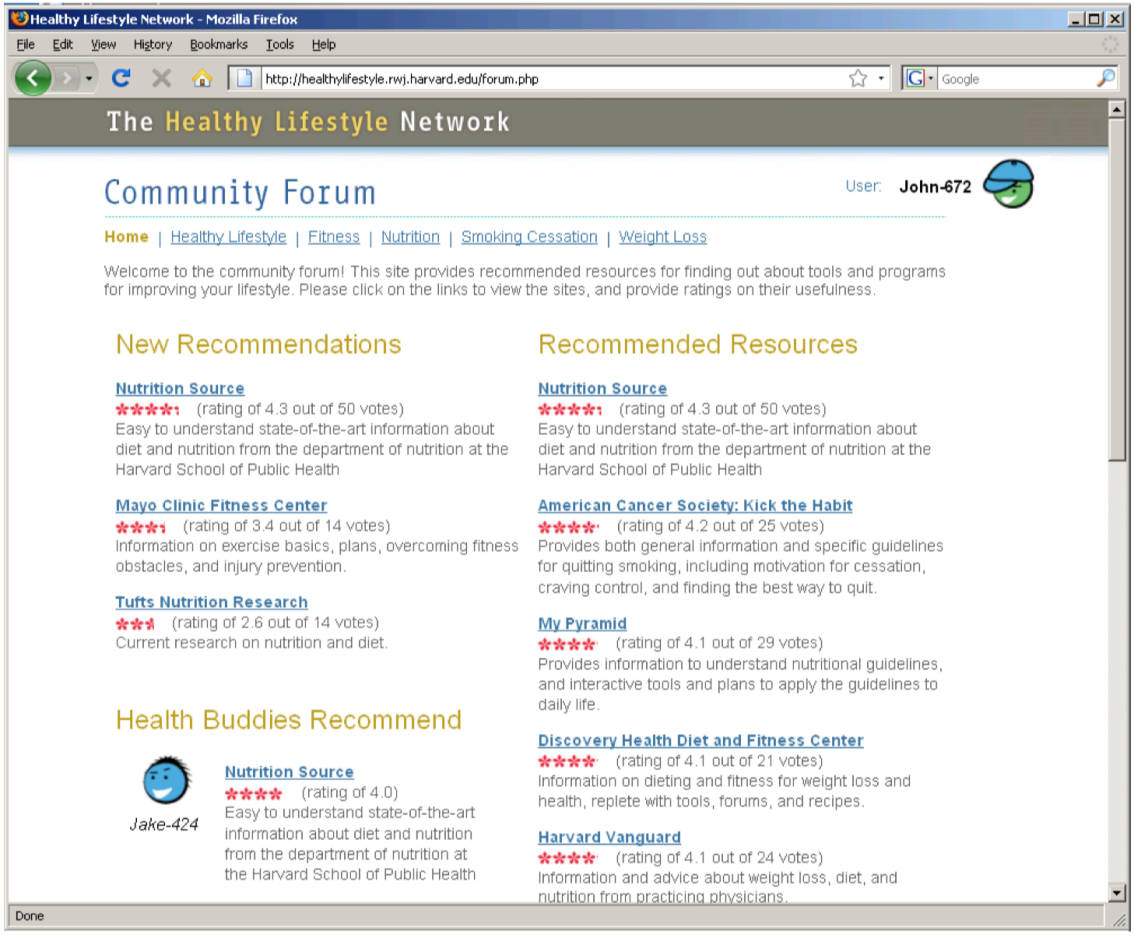


Fig. 1. Randomization of participants to clustered-lattice and random-network conditions in a single trial of this study ($N = 128$, $Z = 6$). In each condition, the black node shows the focal node of a neighborhood to which an individual is being assigned, and the red nodes correspond to that individual's neighbors in the network. In the clustered-lattice network, the red nodes share neighbors with each other, whereas in the random network they do not. White nodes indicate individuals who are not connected to the focal node.

- "The network typologies were created before the participants arrived, and the particpants could not alter the typology in which they were embedded"
  - effectively removing the homophily/selection effect

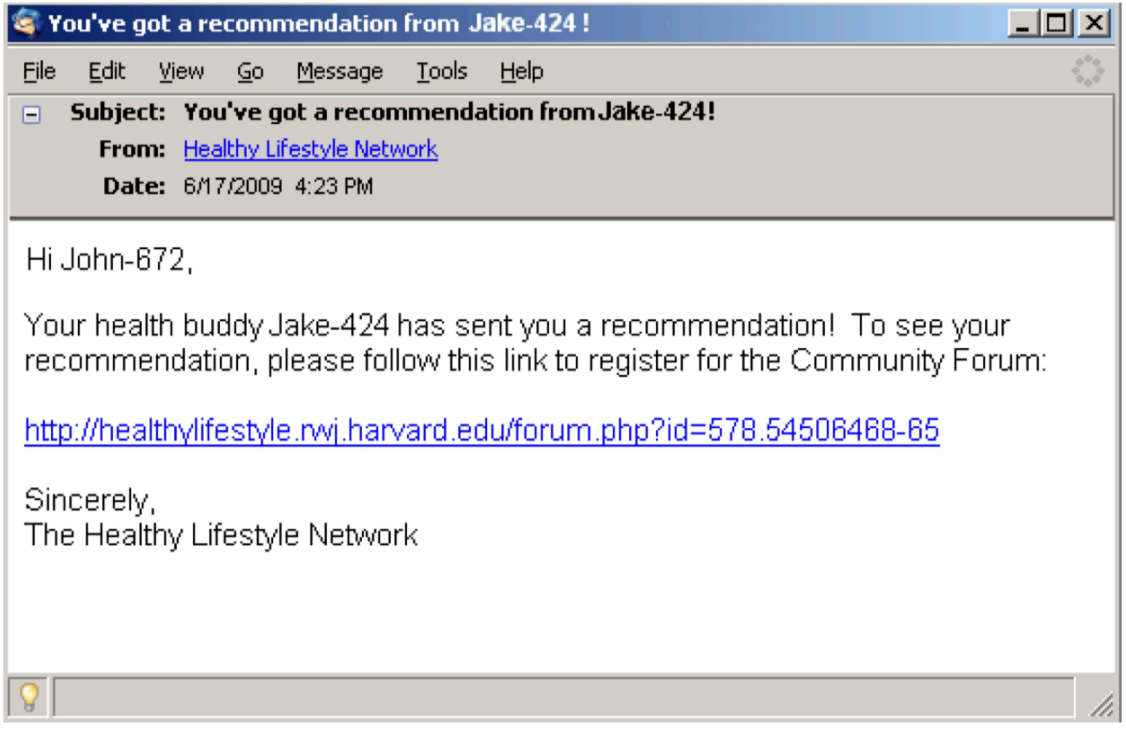
Behavior to diffuse

- Seed nodes will adopt some behaviors first
- Centola think these health behaviors requires multiple confirmation from friends; it's a type of complex contagion



Signals

- Each time a participant adopted the vbehacior, messages were sent to hear health buddies inviting them to adopt



Results

- As expected, contagion is faster on regular networks

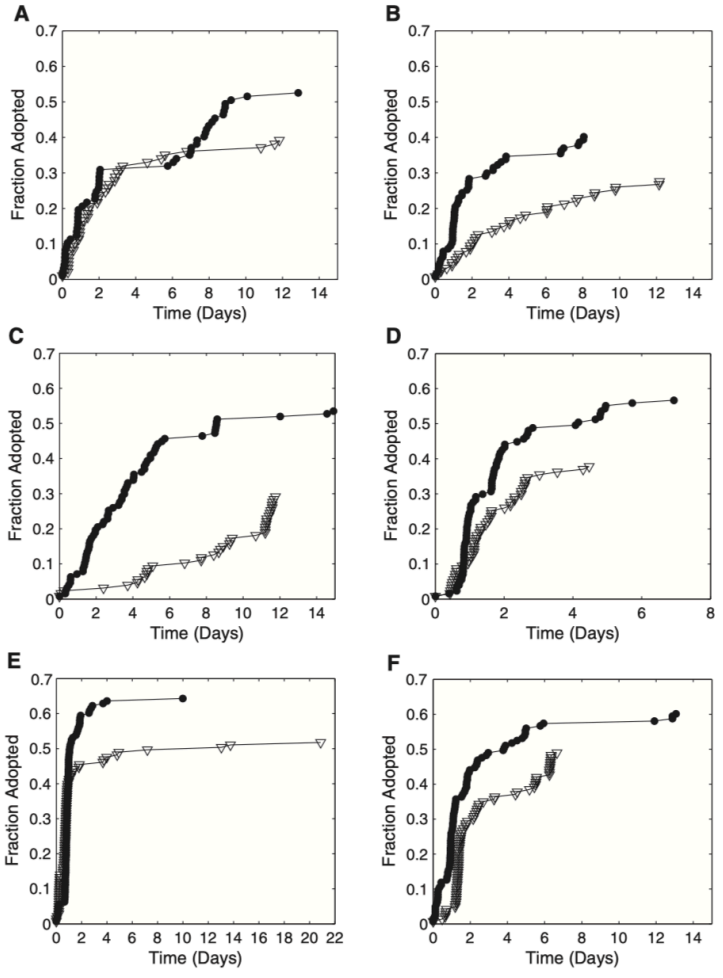


Fig. 2. Time series showing the adoption of a health behavior spreading through clustered-lattice (solid black circles) and random (open triangles) social networks. Six independent trials of the study are shown, including (**A**) $N = 98$, $Z = 6$, (**B** to **D**) $N = 128$, $Z = 6$, and (**E** and **F**) $N = 144$, $Z = 8$. The success of diffusion was measured by the fraction of the total network that adopted the behavior. The speed of the diffusion process was evaluated by comparing the time required for the behavior to spread to the greatest fraction reached by both conditions in each trial.

What to like about this research

- Controls the effects of network topology, independent of factors such as homophily, geographic proximity
- "Study the spread of a health-related behavior that is unknown to the participants before the study, thereby eliminating the effects of nonnetwork factors from the diffusion dynamics, such as advertising, availability, and pricing"
- "allows the same diffusion process to be observed multiple times, under identical structural conditions", thus being more robust
- What are the potential limitations?

Take-away messages

- To make explanatory arguments (e.g., X causes Y)
- Experimental approach usually provides cleaner analysis and more powerful results, compared with statistical analysis on observational data
- But experiments are harder to implement
  - Online field experiment: best if you are insider or know someone in the company
    - Even companies do not want to publish these type of research (though they are still running these experiments everyday)
  - Online lab experiment: Centola spent years to build the website he used

What have we learned?

Lec 2	Prediction (empirical)	Explanation (theory)	Explanation (empirical)
Lec 3	Basic ML		
Lec 4	Text: dictionary		
Lec 5	Text: supervised		
Lec 7	Text: unsupervised		
Lec 8	Text: embedding		
Lec 9		Network: small-world	Network: small-world
Lec 10		Network: weak ties	Network: weak ties
Lec 11			Network: homophily vs contagion

Study Goals

1. Describe the opportunities and challenges of computational social science
2. Evaluate computational social science research on social phenomena
3. Practice the essential techniques to analyze social big data (covered in Tutorials)
    - Getting data
    - Managing data
    - Analyzing data with appropriate methods
4. Propose research questions that are suited to be examined by computational methods with big data
5. Write a research article that utilizes the techniques and methods of computational social science to address social science problems, or design a project that use computational social science to address some real-world problems.

Two types of computational social sciences

- Two parallel developments of computational social sciences
- Studying complex networks
  - Social phenomena are non-linear; we need to study it as a complex network
  - A natural hybrid of theory-driven mathematical simulations and empirical analysis using big data
  - A new paradigm; from studying attributes to studying connections; big mind shift.
- Measurement using Prediction
  - E.g., applying machine learning techniques on text data to generate some variables, and then put these variables into a linear regressions to test some theories
  - Mostly an empirical approach: old theory + new data

Logistics

- Presentation of final project next week
- Let us stick to 15 minutes (including Q&A); so you should prepare slides for 10 minutes
- If you already presented the literature last time; you can skip that part or only briefly mention it
- Think this as an opportunity to solicit feedback and improve your final report;
  - It's not and should not be a finished product

Presentation

Focus on

1. Question
2. What approach you are using? (theory/math modeling, empirical)
3. Research design
   - For empirical work:
     - where is your data? how do you get it? (or have already obtained?)?
     - how do you plan to analyze it?
   - For theoretical work:
     - describe your model
4. Premilinary results