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

Han Zhang

Outline

Friends are similar to each other; why?

Experimental Study of Diffusion Effect

Statistical Identification of Diffusion Effect

Polarization (before social media age)

Polarization (social media age)

Policies of polarization

Summary

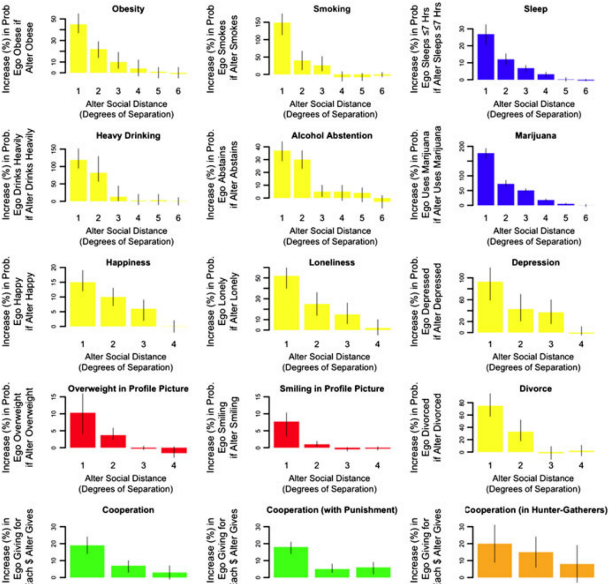
Empirical work on diffusion

- 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

- 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
- 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
- If your goal is to estimate contagion effect, then homophily is essentially an kind of selection bias

Diffusion vs. Homophily

- Diffusion (social influence, contagion) vs. homophily
  - they result in the same correlation pattern: friends are similar to each other
  - but mechanisms are very different
- Social influences and homophily can both explain the below:
  - Friends are similar in their music taste
  - Friends usually consume similar political news (so-called filter bubbles or echo chambers)
  - Smoking/drinking
  - Emotion
  - Political beliefs
- Can you think of other examples?

Alternative mechanism 2: contextual effect

- They just happen 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 statistically
    - get more measures on confounding variables

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 happy messages from friends also make you happier
- Counter-arguments
  - From homophily perspective?
  - From shared context/environment perspective?

Emotion contagion

- 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)
  - Around 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

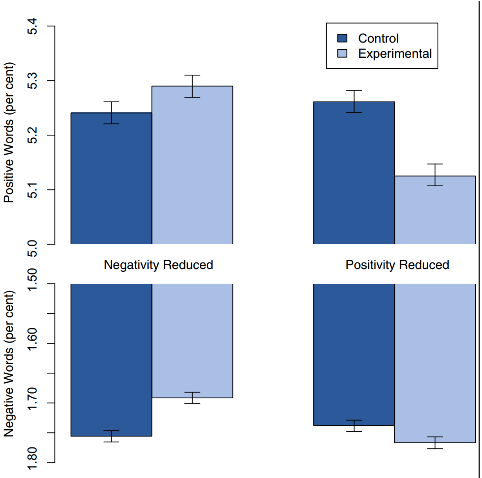


Fig. 1. Mean number of positive (*Upper*) and negative (*Lower*) emotion words (percent) generated people, by condition. Bars represent standard errors.

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 on the ethical debates of Facebook's emotion contagion study

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

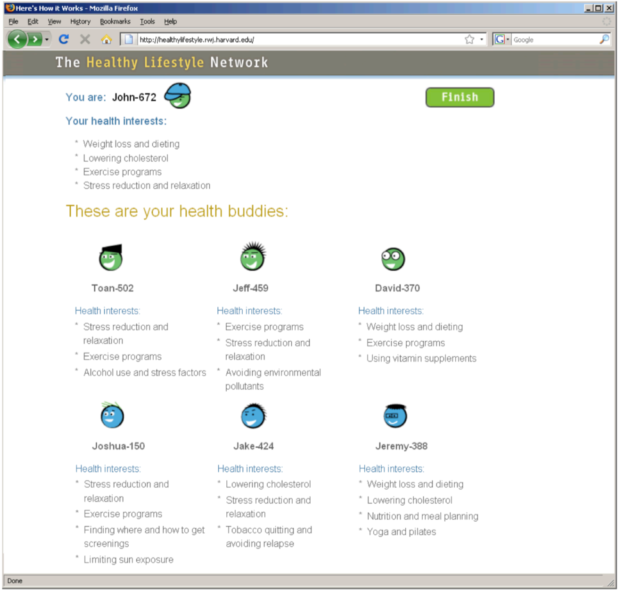
	Regular network (lattice)	Small-world network
Simple contagion	slow	fast
Complex contagion	fast	slow

- 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

Recruitment of research subjects

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

What this health community looks like?



Experiment design

- Some random nodes are reserved for researchers (the seed nodes)
- Respondents were assigned to other nodes and randomly to two conditions

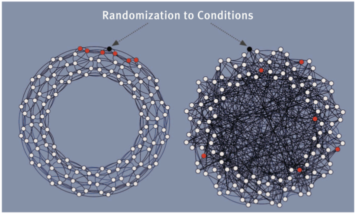
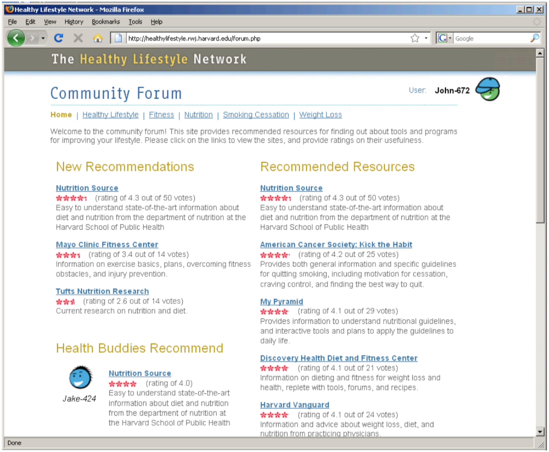


Fig. 1. Randomization of participants to clustered-lattice and random-network conditions in a single trial of this study (N = 128, Z = 6). In each condition, the black node shows the focal node of a neighborhood to which an individual is being assigned, and the red nodes correspond to that individual's 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

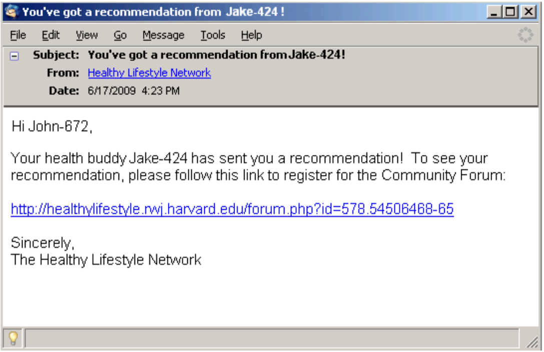
Behavior to diffuse

- 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

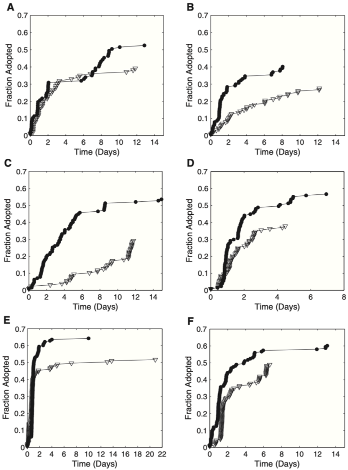


Fig. 2. Time series showing the adoption of a health behavior spreading through clustered-lattice (solid black circles) and random (open triangles) social networks. Six independent trials of the study are shown, including **(A)** $N = 98$, $Z = 6$, **(B to D)** $N = 128$, $Z = 6$, and **(E and F)** $N = 144$, $Z = 8$. 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?

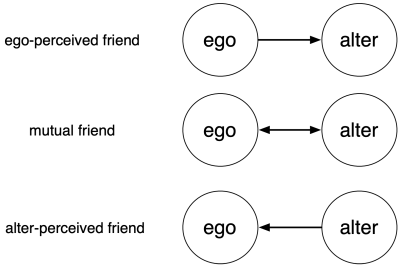
Christakis and Fowler's statistical model

- Christakis and Fowler wants to demonstrate that social influence exists
- They used statistical model to prove that control for homophily and other confounders, social influence still exists
- 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

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

Unobserved factors

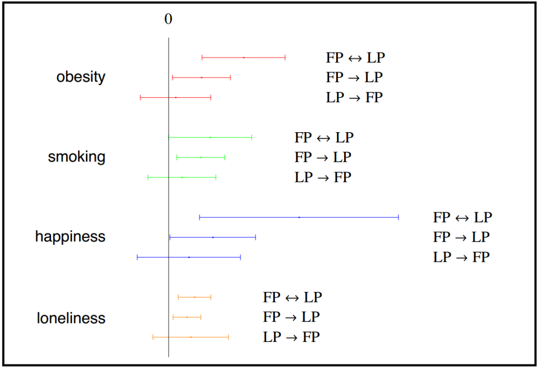
- Even after controlling for many things, we may still have unobserved factors
  - like where do people frequently go, which they have no data on
- To rule out unobserved shared environments: using directionality of ties:



Directionality of ties and unobserved factors

- If social influence is the true mechanism, then
  - mutual friends are most likely to influence each other
  - ego-> alter the second
  - ego<- alter the least likely (you think I am your friends, but I, the survey respondent, do not think you are my friends); So I should not be impacted by you
- If unobserved context matter, then it's likely that you and I share the similar environment regardless of directionality of ties

Directionality of ties and unobserved factors



- FP is ego and LP is alter
- Hence, what we observed is more likely due to social influence, not unobserved contexts

"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)
- They offered a mathematical proof on:

  Contagion effects are nonparametrically unidentifiable in the presence of latent homophily

- Latent homophily are the unobserved similarities that push people to become friends
  - Like genetic similarities. There are arguments that genetically similar people are more likely to make friends with each other
  - But unless your survey have genetics data (mostly likely they don't), you don't have a measure for similarities in genes

An area of debates

- Generally, it is very challenging to separate the three causal mechanisms from observational data
- Christakis and Fowler's data are already better than most because they are panel data and have directionality

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

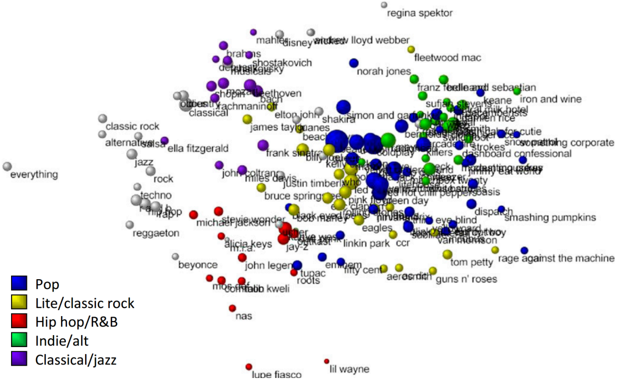
> *Social network analyses of the type employed by Christakis and Fowler will still yield valid tests of the null of no social contagion, even though estimates and confidence intervals may not be valid.*

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)

Diffusion of Music Taste, or no?



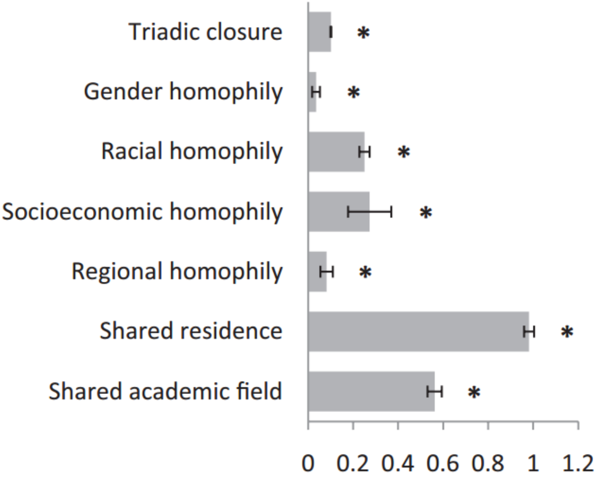
Social influence vs. homophily again

- 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/`

SAOM

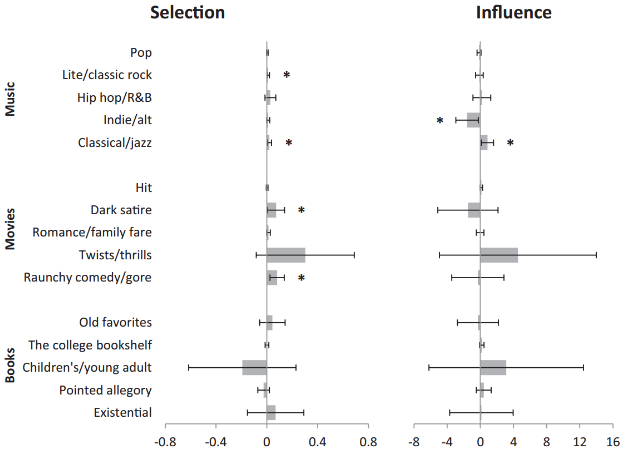
- Simulate what a network would looks like, if both homophily and social influence exists
  - Step 1: fix network structure; allow music taste to diffuse along networks (similar to the examples you have seen last week)
  - Step 2: fix the music tastes; allow people to change their friends, based on similarity of music tastes (homophily)
  - Repeat Step 1 and 2 until convergence
- Now, if we only run step 1 or step 2, then we would know what network looks like if there were no social influence, or no homophily
- Then compare these two simulated networks with observed network

What factors predict friendship



- Shared space is the mostly salient predictor

Homophily vs. social influence



- 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."

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

Take-away messages

- 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
- Both (experimental / observational studies) on estimating causal effects on social networks are frontiers of science right now

Is there polarization?

- Sometimes people worry that too much similarity can be bad
- In politics, this is called polarization
  - friends become too much similar in their political beliefs, and they don't talk/interact with people with other beliefs
- But scientifically, are there any polarization in the first place?

Survey evidence

- Paul DiMaggio, John Evans, and Bethany Bryson, *Have American's Social Attitudes Become More Polarized?*, American Journal of Sociology **102** (1996), no. 3, 690–755
- "Columbia Journalism Review's special "culture wars" issue asserts flatly, "There is increasing polarization in American society"
- One of the early influential empirical study on polarization

Survey evidence: Data

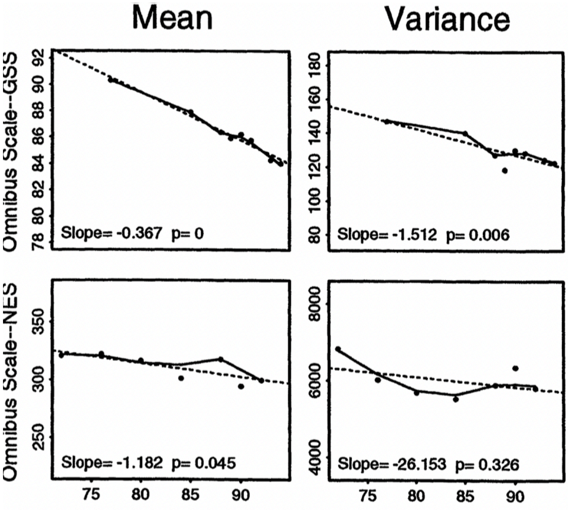
- Context: US
- Data: 20 years of survey data from two nationally Representative surveys, from 1974 through 1994
  - General Social Survey (GSS)
  - National Election Study (NES)
- Survey questions about social attitudes
  - Race: feelings toward black)
  - Poverty: attitudes toward aid to minorities and feelings toward poor people
  - Gender: attitudes toward acceptance of women's occupancy of public roles
  - Crime: attitudes toward crime management (harsh or soft)
  - Abortion

Survey evidence: Methods

- Method:
  - First, select a single question, such as “do you support for more government aids to minorities”
  - Second, calculate variance of the survey responses
    - because variance characterizes the dispersion of
  - Expectation: increasing variance (from surveys) means increasing polarization over the time
- One tricky data manipulation:
  - “don’t know” and “not applicable” -> missing
  - Is this reasonable?

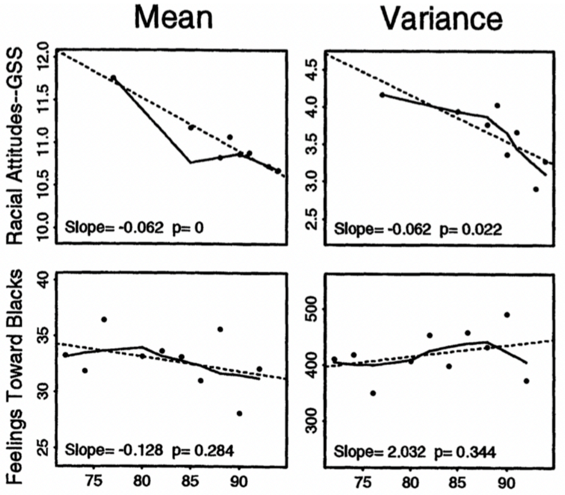
Survey evidence: Results

- Overall polarization (averaged over many different items)



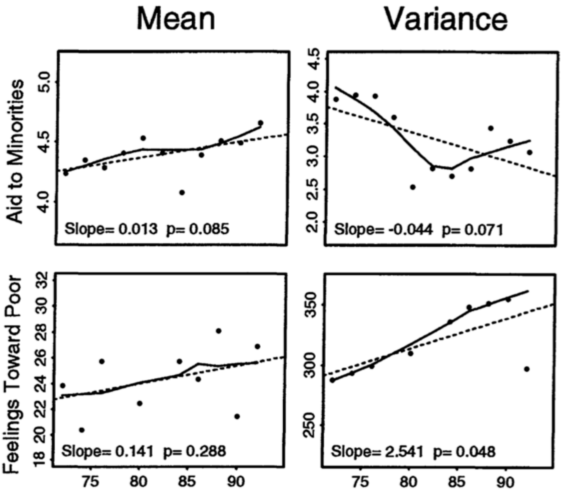
Survey evidence

- Racial attitudes



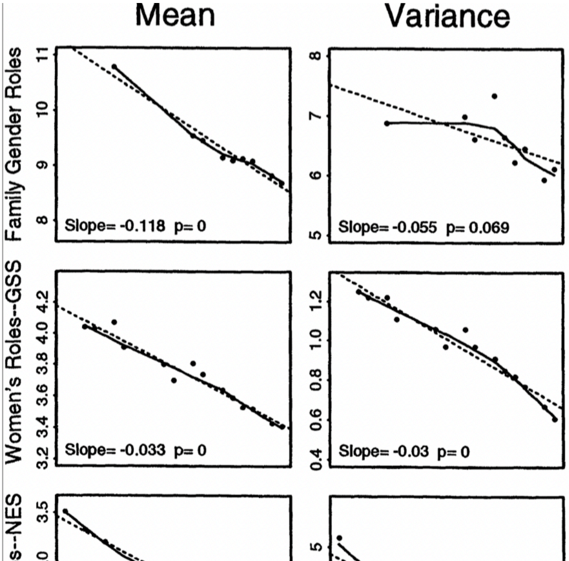
Survey evidence

- Attitudes toward poor



Survey evidence

- Gender norms



Opinion polarization online and offline

- These results were in the last century
- It represents the level of polarization before internet era
- Has internet changed everything?
- Matthew Gentzkow and Jesse M. Shapiro, *Ideological Segregation Online and Offline*, The Quarterly Journal of Economics **126** (2011), no. 4, 1799–1839
- Focus on a particular type of polarization: reading news articles online and offline

Data

- Combines a series of commercial, behavioral, and survey data
- Online news consumption: online website audiences from comScore, supplemented with microdata on the browsing behavior of individuals from 2004 to 2008
  - comScore kept a panel of over 1 million U.S. Internet users. They install software on their computers to permit monitoring of their browsing behavior.
- Offline news consumption: 2008 individual-level data from Mediamark Research and Intelligence on consumption of newspapers, magazines, broadcast television, and cable.
  - 51,354 respondents from the spring 2007 and spring 2008 waves;
  - include frequency of reading certain offline news articles
- Face-to-face interaction: individuals acquaintances and po litical discussants as reported in the 2006 General Social Survey

Method

- Each dataset also includes individual's self-identified political orientation
- Measure ideology of medium: if conservative individuals frequently visit a website, then that website should be more conservative

TABLE II
SIZE AND IDEOLOGICAL COMPOSITION OF ONLINE NEWS OUTLETS

Ten largest

Site	Share of daily visitors who are: Conservative	Liberal	Moderate	Daily UV ('000)
drudgereport.com	.78	.06	.16	475
foxnews.com	.76	.10	.14	1, 159
AOL News	.37	.23	.40	3, 971
usatoday.com	.37	.25	.37	518
msnbc.com	.34	.26	.40	3, 264
Yahoo! News	.31	.25	.43	6, 455
cnn.com	.33	.27	.40	2, 650
nytimes.com	.30	.45	.25	879
huffingtonpost.com	.22	.52	.26	583
BBC News	.16	.57	.26	472

Method

TABLE III

SIZE AND IDEOLOGICAL COMPOSITION OF OFFLINE NEWS OUTLETS

Magazines

	Share of daily readers who are: Conservative	Liberal	Moderate	Market share
Barron's	.43	.19	.37	.02
U.S. News & World Report	.43	.20	.37	.14
BusinessWeek	.42	.21	.37	.07
Forbes	.40	.22	.37	.04
Fortune	.37	.24	.39	.03
TIME	.35	.27	.38	.31
Newsweek	.37	.29	.34	.27
The Economist	.35	.41	.23	.04
The Atlantic	.24	.55	.21	.01
New Yorker	.17	.60	.24	.07

Method

- Individual-level
  - we have a measure for whether a site is more conservative or not
  - If one individual further visits a lot of conservative websites
  - Then he is really conservative
- Isolation index: average conservative exposure of conservative contents minus the average conservative exposure of liberal contents
- Isolation index $= 0$: all conservative and liberal visits are to the same outlet; no polarization
- Isolation index $= 1$: conservatives only visit 100% conservative outlets and liberals only visit 100% liberal outlets; complete polarization

Results

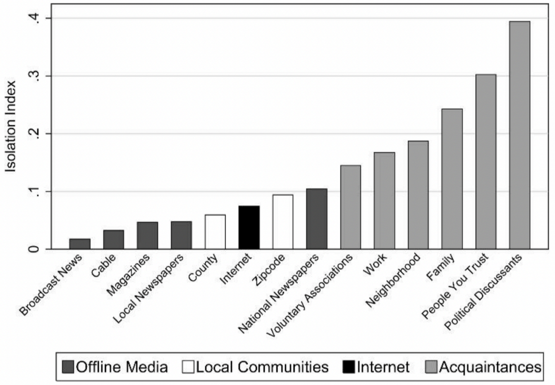


FIGURE II

Results

- Somewhat surprising
- Is there polarization in reading online political news? Yes
- Is the online polarization high? no; much lower than non-internet mediums

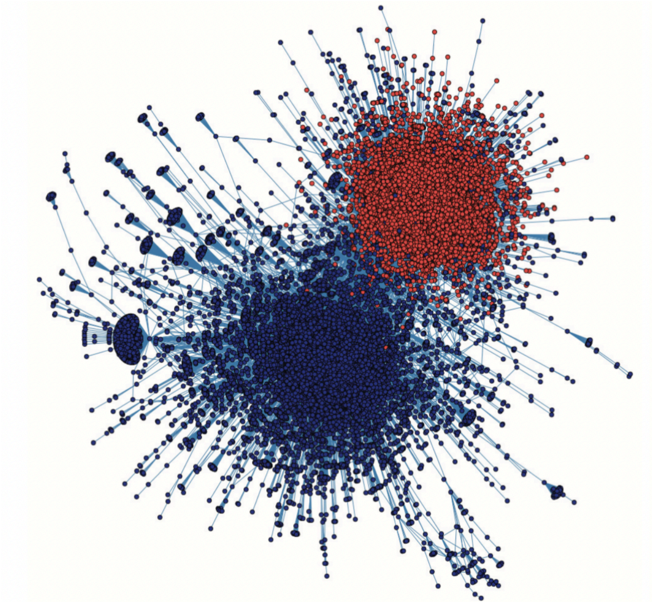
Polarization on Social Media

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini, *Political Polarization on Twitter*, Proceedings of the International AAAI Conference on Web and Social Media **5** (2011), no. 1, 89–96

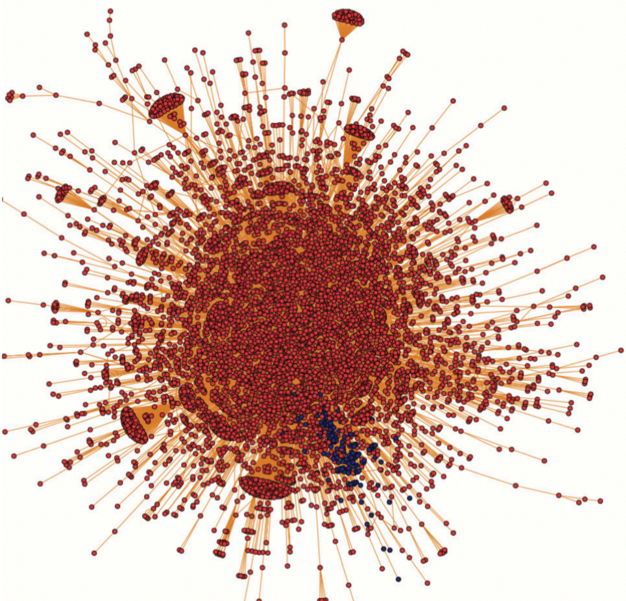
Results

- Manually curated 66 politics-related hashtags on Twitter
- Found 252,300 Twitter posts with these hashtags
- Belonged to 45,365 users
- 23,766 had retweeted at once
- The largest connected component contains 18,470 nodes
- Methods: network analysis + text analysis

Results: retweet networks



Results: mention networks



Comparisons

- survey
- behavioral data from commercial companies
- social media
  - automatically have text + network + images and many more
- The same phenomenon can be studied in different types of data and using different methods

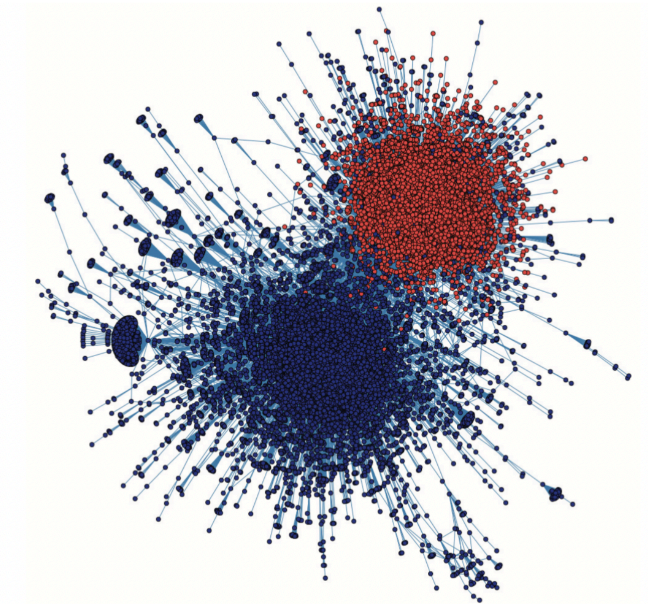
Polarization on Social Media

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini, *Political Polarization on Twitter*, Proceedings of the International AAAI Conference on Web and Social Media **5** (2011), no. 1, 89–96

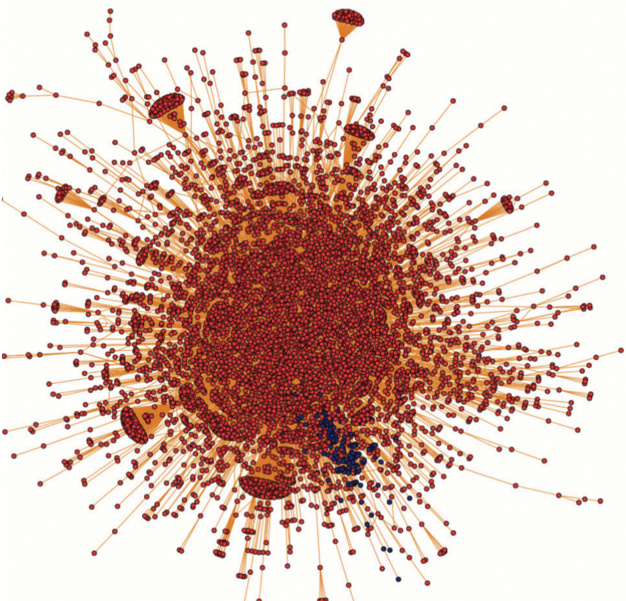
Results

- Manually curated 66 politics-related hashtags on Twitter
- Found 252,300 Twitter posts with these hashtags
- Belonged to 45,365 users
- 23,766 had retweeted at once
- The largest connected component contains 18,470 nodes
- Methods: network analysis + text analysis

Results: retweet networks



Results: mention networks



Mechanisms and policy implications

- Say, you observe that friends' political beliefs are very similar.
- Social influence: encourage them to talk more about non-political contents?
- Homophily: let Facebook recommend some friends from the other political view in the beginning
- Context: try to downgrade the influence of some common information sources? Like ban some common low-quality information sources?
- As you can see, if we know which mechanism has more impact, we can design better policies

Causes of polarization

- The next question to ask is what causes polarization
- There is a popular belief that Internet and social media have fostered polarization, because there is echo chamber effects:
- For instance, Cass Sunstein, in his 2001 book *Republic.com*, argues that Internet prevent people from learning about opinions from other sides, which ultimately weakens the basis for democracy
  - Sunstein made popular the idea of echo chambers (also called filter bubbles)
- Based on our previous study, what's a competing explanation of political polarization on internet and social media?

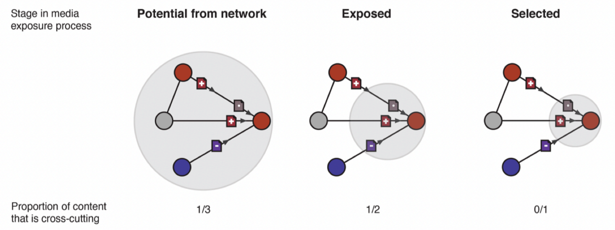
Causes of polarization:

E. Bakshy, S. Messing, and L. A. Adamic, *Exposure To Ideologically Diverse News And Opinion On Facebook*, Science **348** (2015), no. 6239, 1130–1132

- Social influence: Friends share content that align with your political ideology , which creates echo chamber)
- Homophily (self-selection):
  - strong version: you don't want to be friend with different political views
  - weak version: even if you have that type of friends, if they share opposite-view links, you are not going to read them; you will only read news aligned with your beliefs
- Specific to this study: Facebook's algorithm impact (algorithm confounding)

Methods

- Three types of exposure, from July 2014 to Jan 2015
- 3.8B potential exposure: your friends shared news; may not appear on your News Feed
- 903M exposure: content actually appeared on your screen
- 59M clicks: you actually clicked the shared content

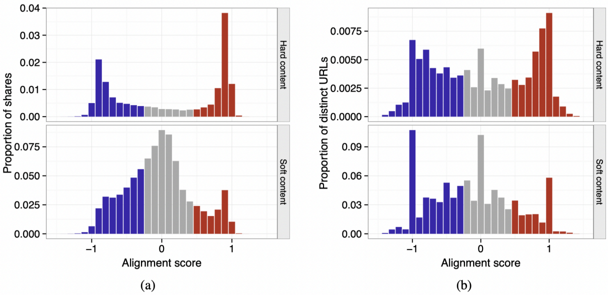


Methods: data cleaning and generation

- Remove irrelevant ones
  - Friends may share all sorts of contents; they care about political contents (they call "hard news")
  - Tagged around 150,000 stories as training data
  - Further trained a SVM classifier, distinguish hard and soft news
  - around 9% are classified as hard, political news; the rest are soft news
- Generate political ideology (they call this alignment score)
  - Similar to Gentzkow and Shapiro, they discarded every Facebook user who did not report their political ideology
  - Then if a website was clicked by more conservative users, it is more conservative
  - +1 means conservative ; -1 means liberal

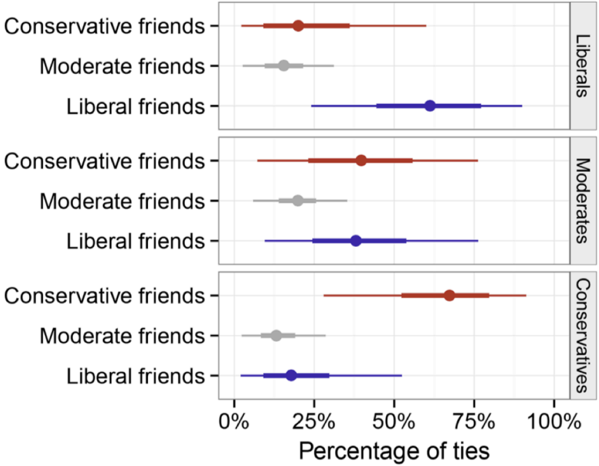
Results

- Mostly shared links on political news aligned with liberal or conservative subpopulations
- Not so for soft news



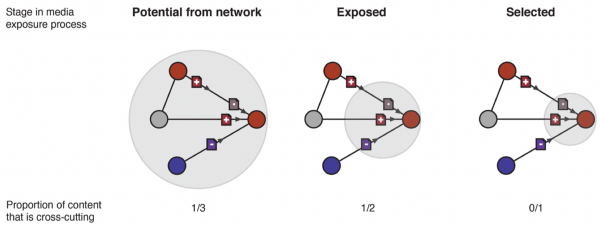
Results: similarities of ideology between friends

- Friends are clearly similar to each other in political ideology
- But there are many cross-cutting ties



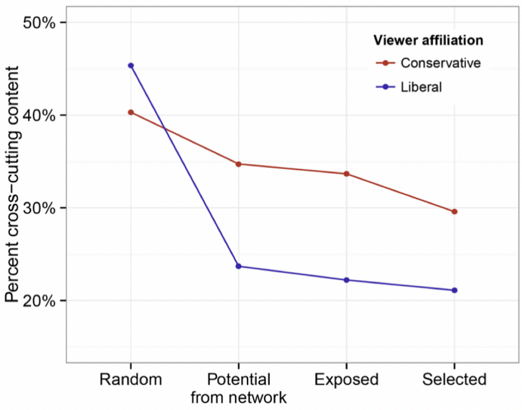
Results: effect of Facebook's algorithm confounding

- These authors were at Facebook so they can evaluate how much more content from the same camp was resulted by Facebook's ranking algorithm for News Feed
- "conservatives see approximately 5% less cross-cutting content compared to what friends share"
- "while liberals see about 8% less ideologically diverse content"
- So they think it's not a large impact



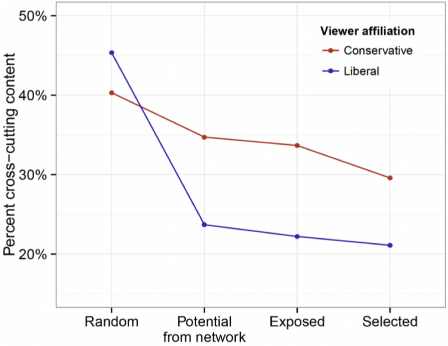
Results: homophily

- Instead, they argue that homophily plays a large role (people choosing content aligned with their ideology)
  - Particularly for conservatives, decrease by about 17%
  - 6% for liberals



Results: echo chambers

- Still, substantive room for every user to be exposed to, and to actually click on cross-cutting contents
- Is this high or low? Without a rigorous study of other platforms, it's hard to know
  - That's why Gentzkow and Shapiro's study is so valuable



Compare three mechanisms

- Social influence/diffusion/contagion
- Homophily
- They all result in same correlation, but policy implications are very different

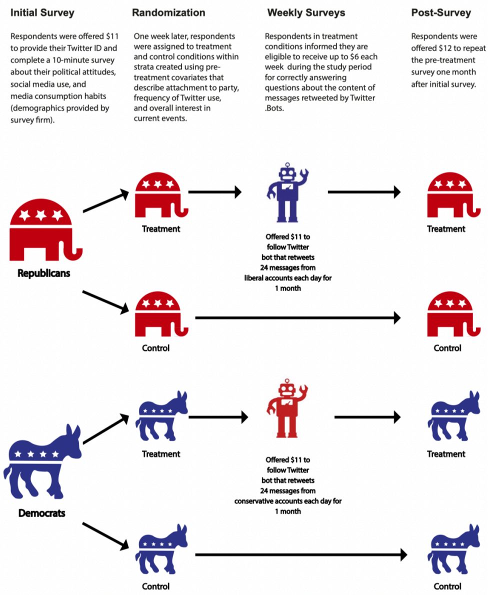
Policy implications

- How can you design politics to reduce similarity in friends' political beliefs?
- If homophily were the main cause?
  - harder; you would prevent people adding ties to friends similar to them
  - or prevent people from read like-minded contents shared from friends
  - companies may not like this idea
- if social influence were the main cause?
  - expose more cross-cutting contents to social media users
  - This is more doable; can be embedded in recommendation algorithms

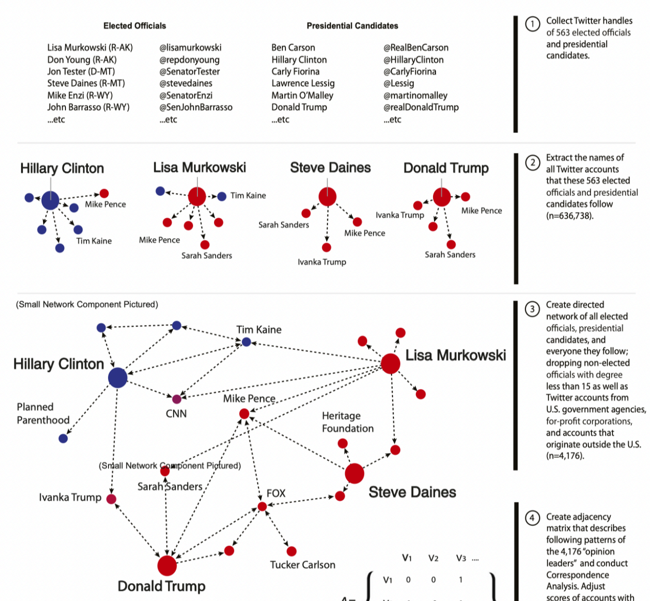
Experiment; setup

- If a company adopt such policy, would it really work?
  - We need to evaluate the impact of this policy.
  - Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky, *Exposure to opposing views on social media can increase political polarization*, Proceedings of the National Academy of Sciences **115** (2018), no. 37, 9216–9221
- Bail is not affiliated with any social media company; so he could not access to full social media data, unlike Bakshy et al.
- Without access to company's data, if they were to perform a study using only observational data, people would still challenge them questions about causality
- So they designed a randomized controlled experiment
  - 901 Democrats and 751 Republicans recruited

Experiment design



Measure ideology



Measure ideology

- Then apply a kind of dimensional reduction to obtain ideology scores for opinion leaders
- Finally, random sample real tweets published by these opinion leaders, and feed them to subjects

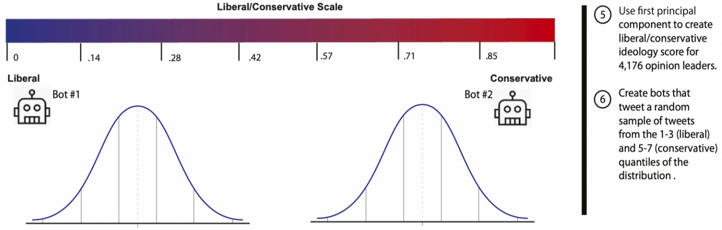


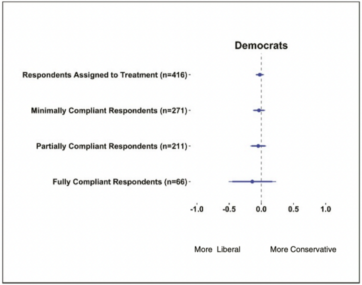
Fig. 2. Design of study's Twitter bots.

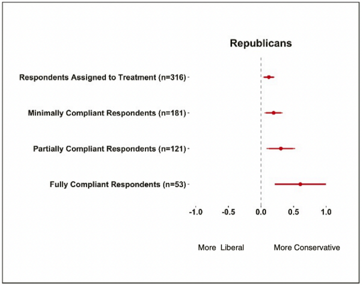
Measure outcomes

- Treatment compliance:
  - Treated users were asked to follow the bots, but they may not read the contents
  - If someone is assigned to treatment group but did not follow researcher's instruction, it's called noncompliance
  - To measure compliance, participants were offered additional $18 to complete weekly surveys that ask them questions about content of the tweets produced by the bots
- Outcome:
  - There is a pre-treatment survey
  - Respondents were asked to fill a final survey, which asked them the same set of questions
  - The study period is about a month

Results

- Exposure to counter-attitudinal information backfire, actually increased polarization





Today

- Phenomena: friends are similar
- Two different contexts: diffusion; polarization and echo chamber
- Causes: homophily or social influences?
- Empirical approaches to causal inference problem: experimental vs statistical

What have we learned?

Lec 2&3	Prediction (empirical)	basics
Lec 4-7	text analysis	measurement (empirical)
Lec 8-10	network analysis	Explanation (theory + empirical)

- Two parallel developments of computational social sciences
- Studying complex networks
  - 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