The Structure of Social Data

http://www.stanford.edu/~jugander/mse334/

Johan Ugander
Assistant Professor, Management Science & Engineering
Stanford University
9/22/2015

Course outline

Week 1: Introduction and Overview

Week 2: Heavy-tailed distributions; Random graphs

Week 3: More graph models

Week 4: Graph centrality and ranking

Week 5: Ranking from comparisons; Friendship paradoxes

Week 6: Homophily, social influence, contagion

Week 7: Causal inference on networks

Week 8-9: Synthesis; dissecting complex papers

Week 10: Project presentations

Practical components

Problem Set 1: Manipulating data (due 10/8)

Problem Set 2: Theory exercises (due 10/20)

Practical components

Problem Set 1: Manipulating data (due 10/8)

Problem Set 2: Theory exercises (due 10/20)

Reaction paper: (due 10/27)

Groups of 1–3. Skim some papers on the course homepage. Pick a paper to read carefully; pick a citation there–in, read it carefully.

Prompt: What are the weaknesses of the papers, and how could they be improved? What are some promising further research questions in the direction of the papers, and how could they be pursued?

10/27–11/3: Discussion meetings with me!

1 week after meeting: Project proposal

Projects

Basic genres:

– An empirical evaluation of an algorithm, model, or measure on an interesting dataset.

– A theoretical project that considers an algorithm, model, or measure in the area of some course topic, and derives rigorous results about it.

– An extended, critical survey of one the course topics, going into significant depth and offering a novel perspective on the area.

12/1:	**Report due**
12/1 & 12/3:	**In-class presentations**

Questions?

Learning outcomes

1. Students should develop a familiarity with relevant structural properties of empirical social networks, and how different graph models capture or don't capture these properties.

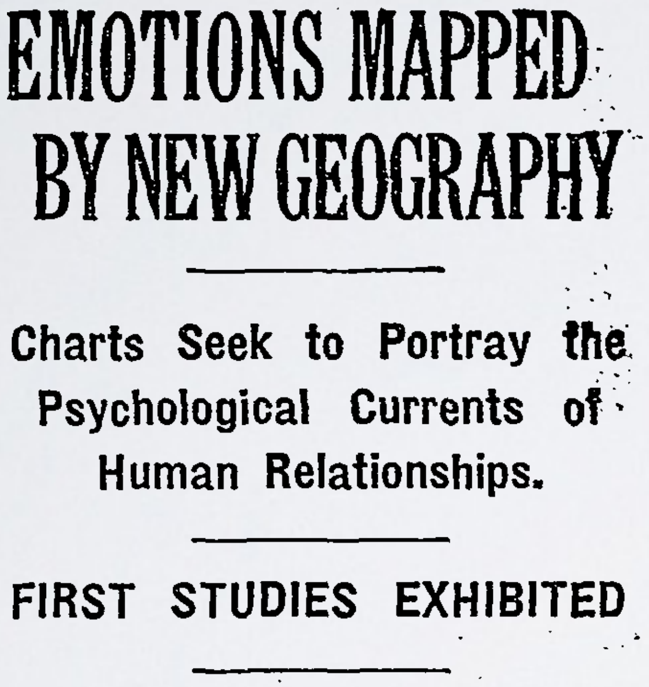
2. Students should be able to weigh advantages and disadvantages of different observational and experimental study designs that examine/test mechanisms in social systems.

3. Students should be able to employ structural measures for diverse ranking/predicting problems on graphs.

4. Students should be able to critically read research papers in the field to identify strengths and potential weaknesses, and to be able to design tests of potential weaknesses.

Ready?

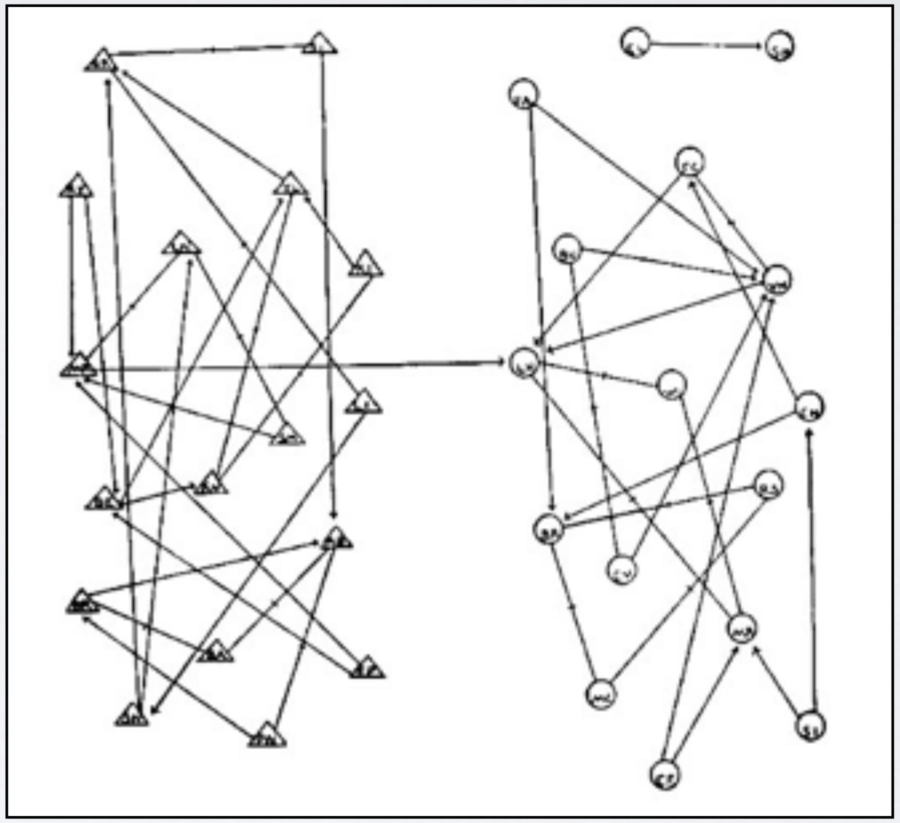
Social networks: mapping structure



EMOTIONS MAPPED BY NEW GEOGRAPHY

Charts Seek to Portray the Psychological Currents of Human Relationships.

FIRST STUDIES EXHIBITED

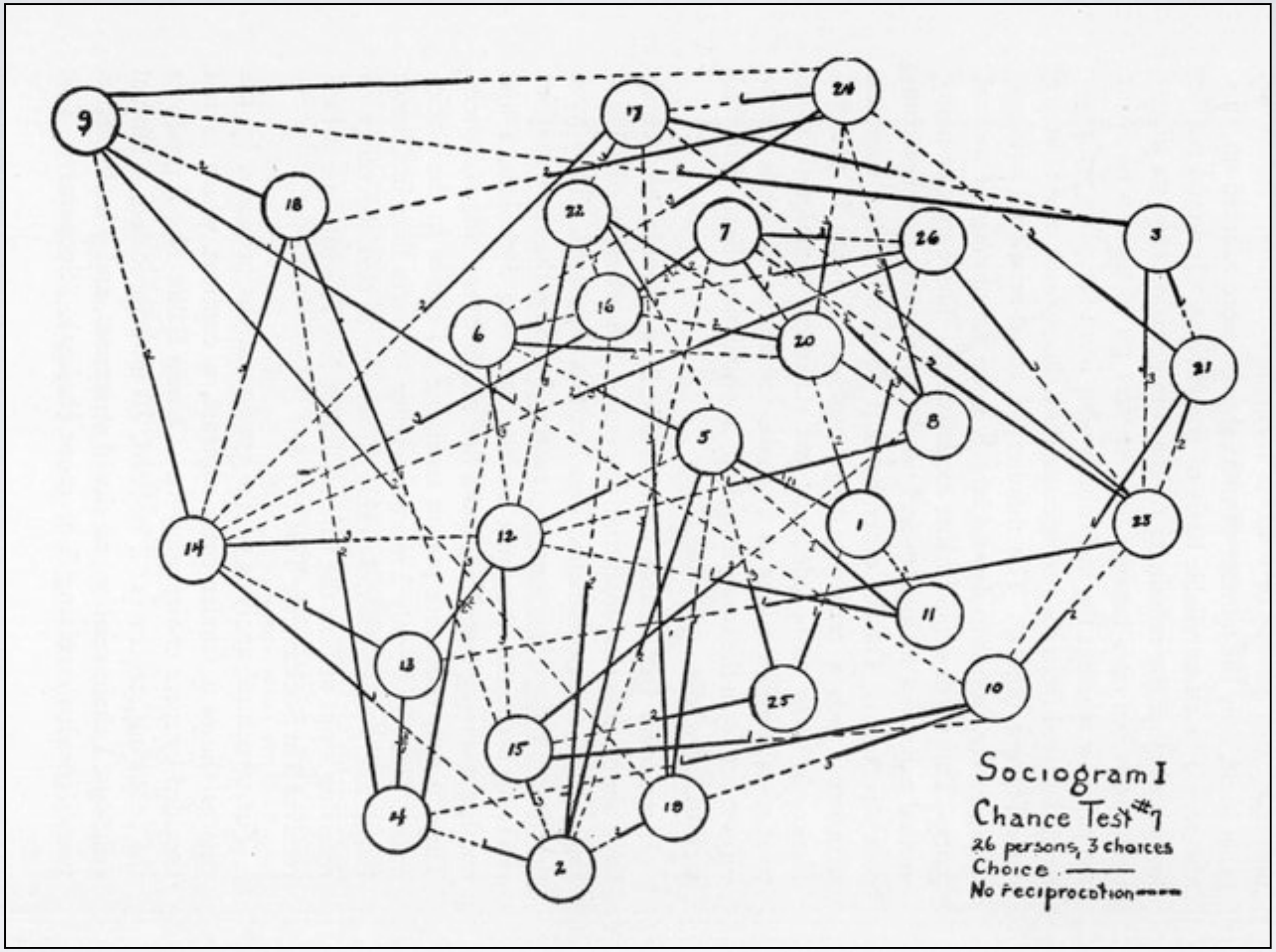


n=33

First “sociogram”: 8th grade students studying in proximity

- J Moreno (1934) “Who shall survive?: A new approach to the problem of human interrelations.”

Social networks: mapping structure



Moreno's “chance sociogram”: a random graph null model

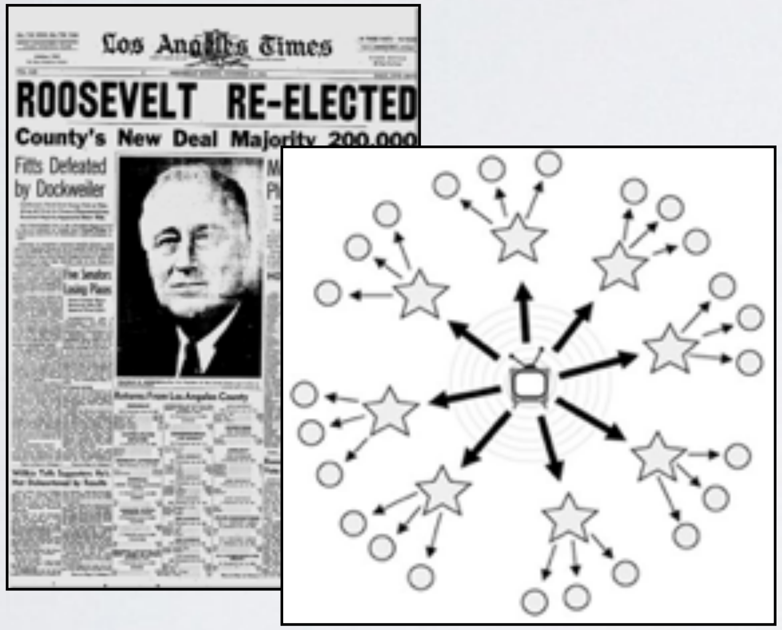
- J Moreno (1934) “Who shall survive?: A new approach to the problem of human interrelations.”

The digital microscope

n>1,400,000,000

Processes on social networks

1940 election: two-step theory of opinion leaders



Los Angeles Times

ROOSEVELT RE-ELECTED

County's New Deal Majority 200,000

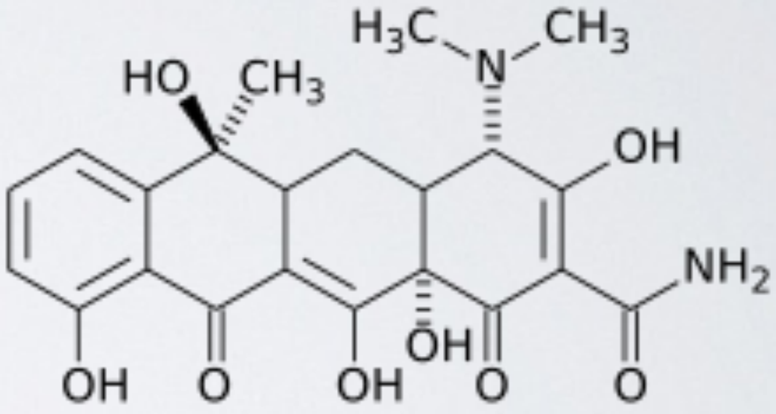
Fitts Defeated by Dockweiler

Lazarsfeld et al. '55
Watts-Dodds '07

Hybrid seed corn

Ryan-Gross '43

Tetracycline



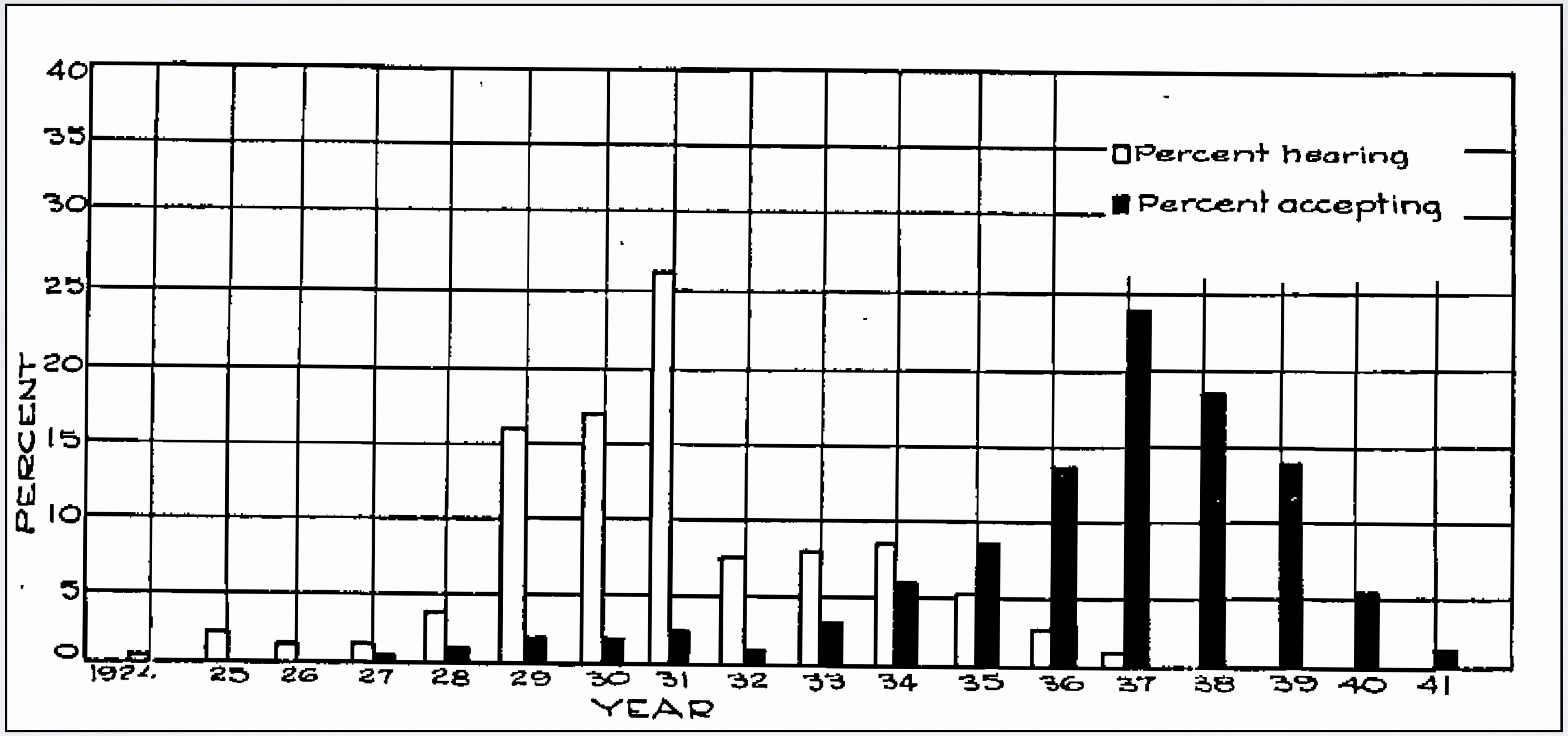
Coleman-Katz-Menzel '57

- B Ryan, N Gross (1943) "The diffusion of hybrid seed corn in two Iowa communities", Rural sociology.
- P Lazarsfeld; B Berelson, H Gaudet (1948) "The People's Choice. How the Voter Makes up His Mind in a Presidential Campaign".
- E Katz, P Lazarsfeld (1955) "Personal Influence, The part played by people in the flow of mass communications".
- E Katz (1957) "The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis". Political Opinion Quarterly.
- J Coleman, E Katz, H Menzel (1957) "The diffusion of an innovation among physicians", Sociometry.
- D Watts, P Dodds (2007) "Influentials, Networks, and Public Opinion Formation" Journal of Consumer Research.

Processes on social networks

Hybrid seed corn (Ryan-Gross):

5 stages: awareness, interest, evaluation, trial, adoption

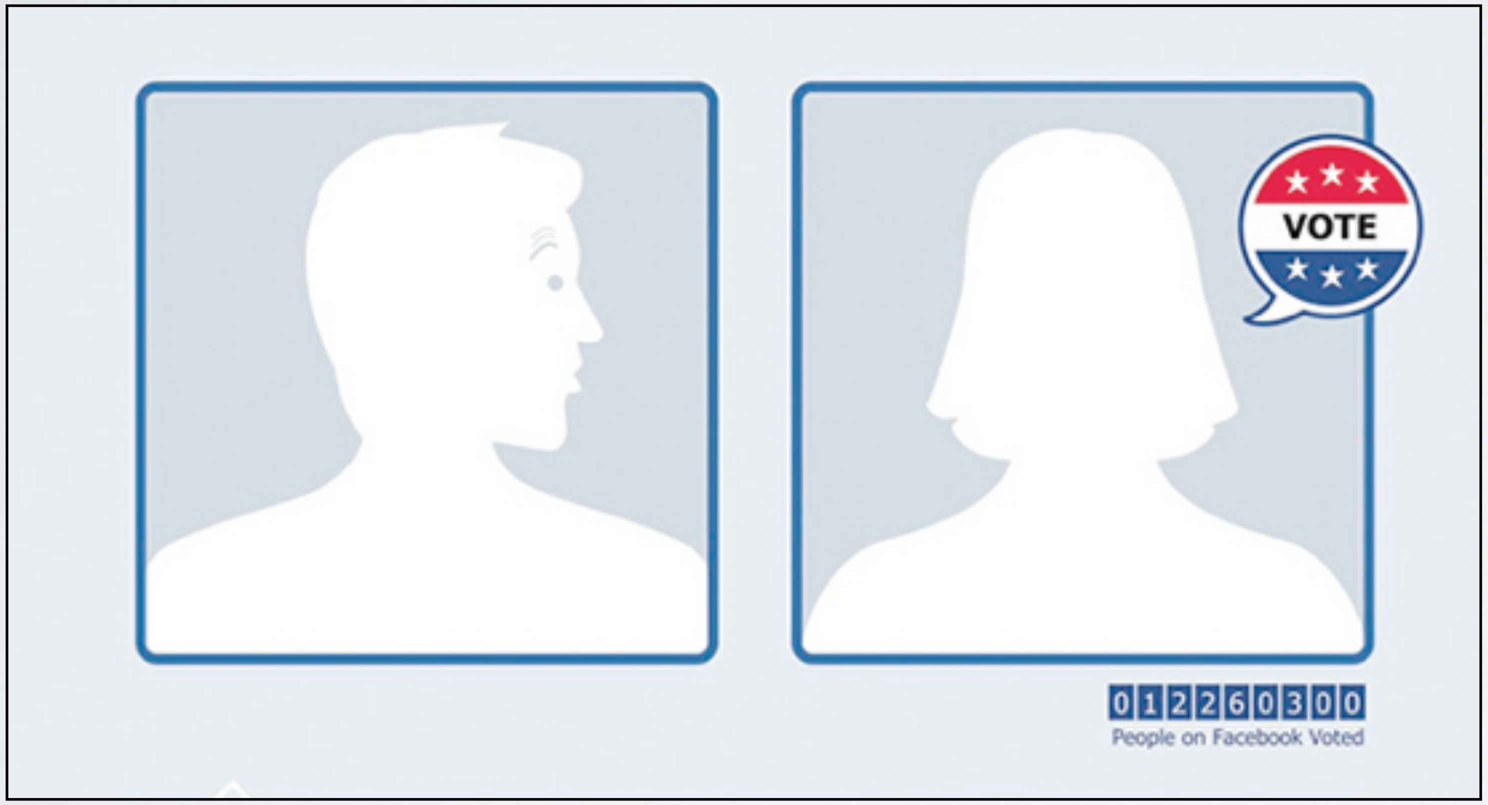


Survey of n=259 farmers

- B Ryan, N Gross (1943) “The diffusion of hybrid seed corn in two Iowa communities”, Rural sociology.
- P Lazarsfeld; B Berelson, H Gaudet (1948) “The People's Choice. How the Voter Makes up His Mind in a Presidential Campaign”.
- E Katz, P Lazarsfeld (1955) “Personal Influence, The part played by people in the flow of mass communications”.
- E Katz (1957) “The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis”. Political Opinion Quarterly.
- J Coleman, E Katz, H Menzel (1957) “The diffusion of an innovation among physicians”, Sociometry.
- D Watts, P Dodds (2007) “Influentials, Networks, and Public Opinion Formation” Journal of Consumer Research.

Digital experimental microscope

Massive experiments to test theories of social processes on large-scale networks.



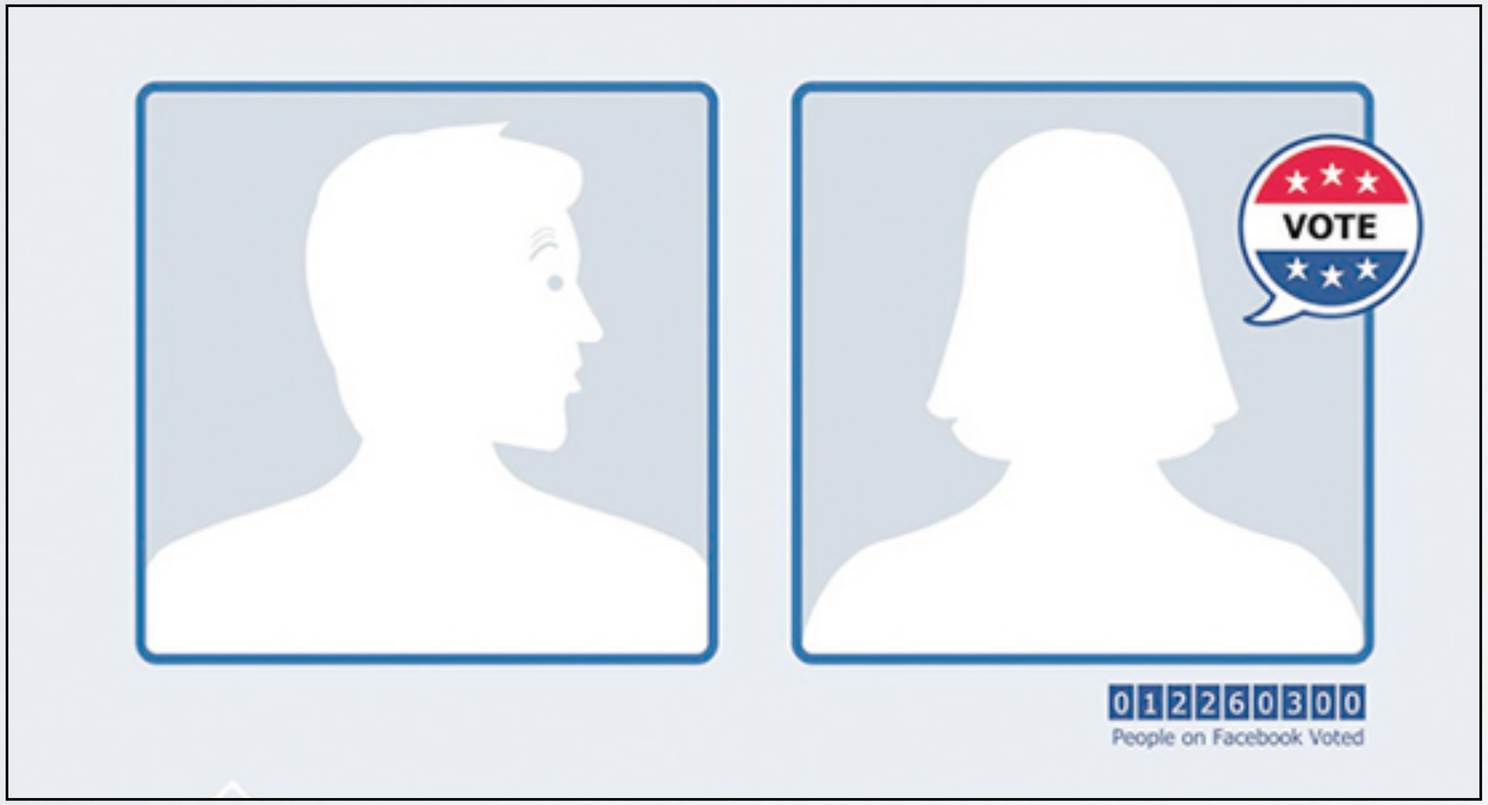
Experiment on n=61,000,000 Facebook users

Not just FB: Telenor service experiment (n=46,000), LinkedIn, others.

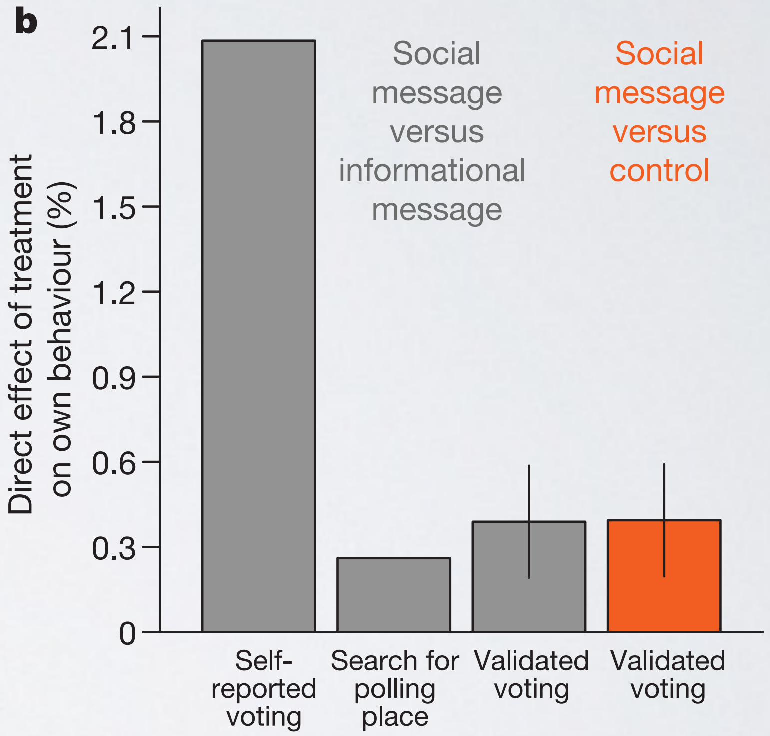
- Bond et al. (2012) “A 61-Million-Person Experiment in Social Influence and Political Mobilization”, Nature.
- J Bjelland et al. (2015) “Investigating Social Influence Through Large-Scale Field Experimentation”, NetMob.

Digital experimental microscope

Massive experiments to test theories of social processes on large-scale networks.



Experiment on n=61,000,000 Facebook users

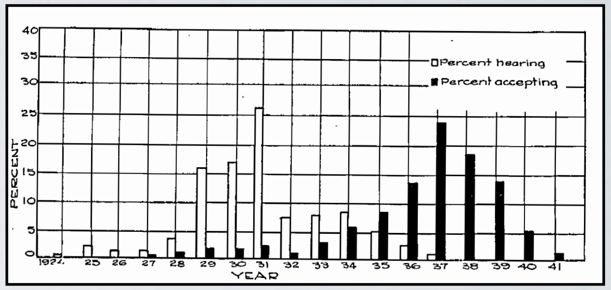


Not just FB: Telenor service experiment (n=46,000), LinkedIn, others.

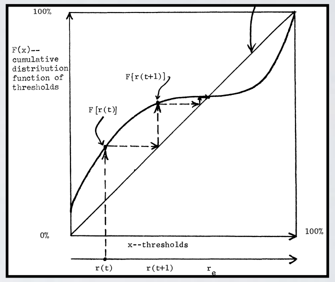
- Bond et al. (2012) “A 61-Million-Person Experiment in Social Influence and Political Mobilization”, Nature.
- J Bjelland et al. (2015) “Investigating Social Influence Through Large-Scale Field Experimentation”, NetMob.

Timeline

- 1940s–50s: Early theories, early data

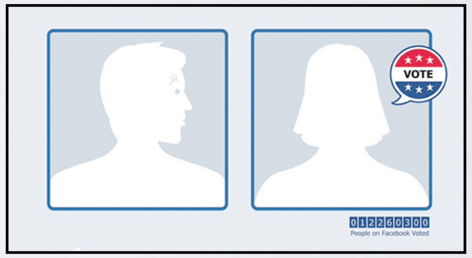


- 1960s–90s: Theory refinement/testing



- 2000s: Large-scale data

- 2010s: Large-scale experiments



Designing/analyzing experiments to develop/test network theories

= Big opportunity

Whirlwind tour

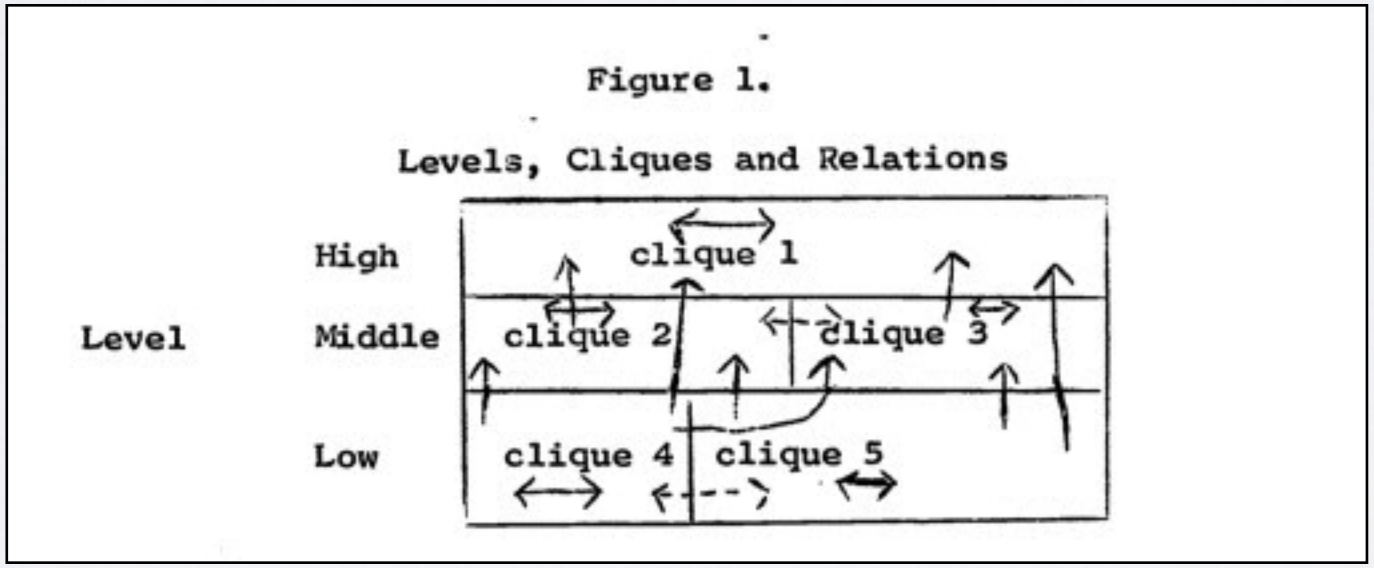
- What network?
- From karate to communities
- Influence, instrumented
- Homophily vs. contagion
- Influence experiments
- Network experiments

Hopefully this lecture will convince you there are really interesting research questions you can work on.

Testing a theory with data

- Homans (1950): Small groups of people create a social structure that contains many clique subgroups and a ranking system.

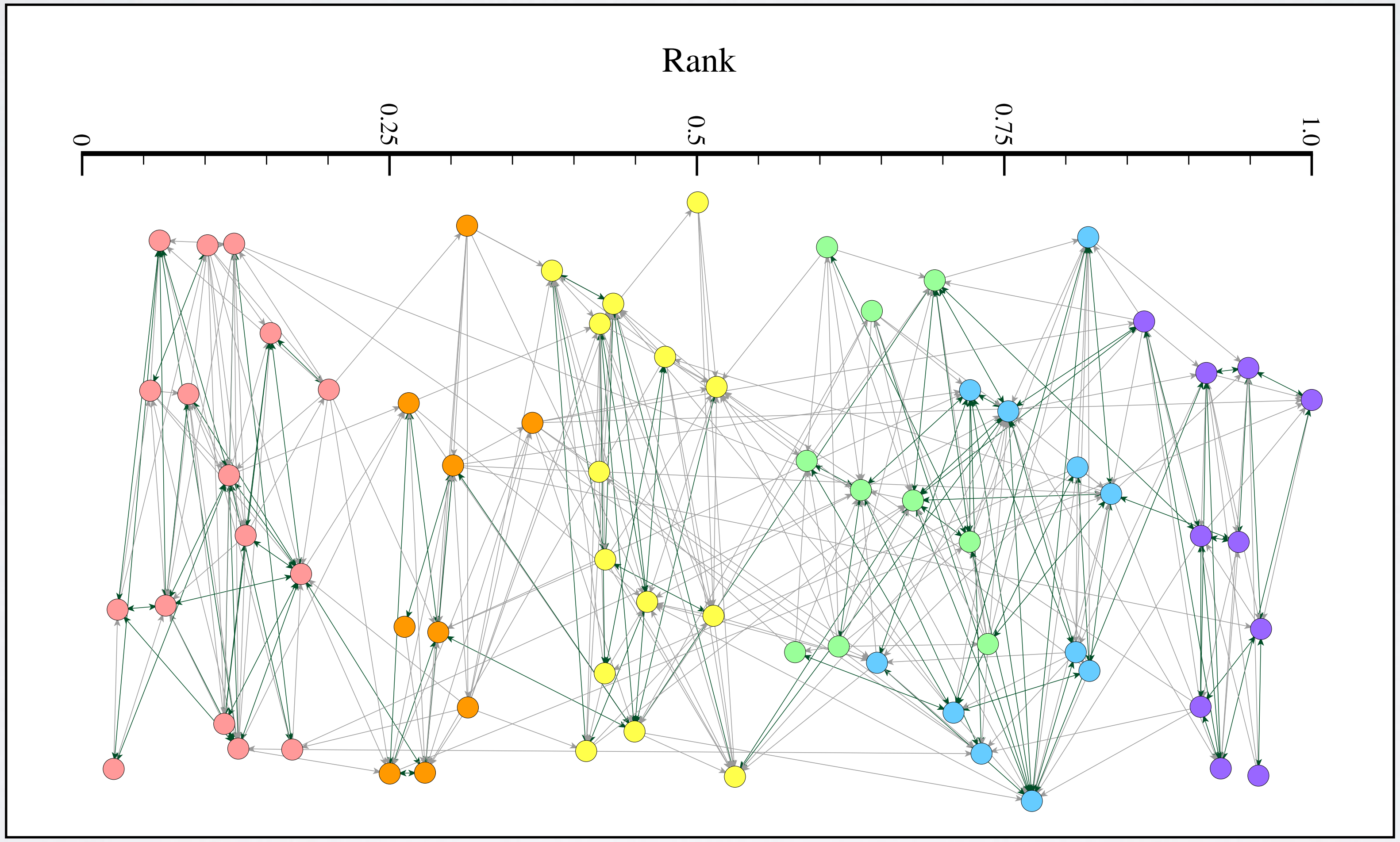
- Davis & Leinhardt (1967): Operational statement using subgraph frequencies: some patterns less frequent than random model predicts.



- G Homans (1950) "The Human Group."
- J Davis, S Leinhardt (1971) "The structure of positive interpersonal relations in small groups," Sociological Theories in Progress.

Extending a theory with more data

- Ball–Newman (2013): Maximum likelihood inference of status according to Homans' theory.
- Examined 84 high school networks for correlates of Homans status.



- G Homans (1950) "The Human Group."
- J Davis, S Leinhardt (1971) "The structure of positive interpersonal relations in small groups," Sociological Theories in Progress.
- B Ball, MEJ Newman (2013) "Friendship networks and social status." Network Science.

What network?

“Name generators” in sociology show huge difference between social networks generated by questions:

“Who do you know?” **“Who are your three closest friends?"**
“With whom do you discuss important matters?”

- K Campbell, B Lee (1991) “Name generators in surveys of personal networks.” Social Networks.
- M Resnick et al. (1997) “Protecting adolescents from harm: findings from the national longitudinal study on adolescent health.” JAMA.
- C Apicella, F Marlowe, J Fowler, N Christakis (2012) “Social networks and cooperation in hunter-gatherers,” Nature.
- B Ball, MEJ Newman (2013) “Friendship networks and social status.” Network Science.

What network?

“Name generators” in sociology show huge difference between social networks generated by questions:

“Who do you know?” **“Who are your three closest friends?"**
“With whom do you discuss important matters?”

Ball–Newman used AddHealth, which has issues that requires care.

“Up to 10 people with whom they are friends,
with a maximum of five being male and 5 being female”

- K Campbell, B Lee (1991) “Name generators in surveys of personal networks.” Social Networks.
- M Resnick et al. (1997) “Protecting adolescents from harm: findings from the national longitudinal study on adolescent health.” JAMA.
- C Apicella, F Marlowe, J Fowler, N Christakis (2012) “Social networks and cooperation in hunter–gatherers,” Nature.
- B Ball, MEJ Newman (2013) “Friendship networks and social status.” Network Science.

What network?

“Name generators” in sociology show huge difference between social networks generated by questions:

“Who do you know?” **“Who are your three closest friends?"**
“With whom do you discuss important matters?”

“With whom they would like to live in the next camp”

“To whom they would give an actual gift of honey”

- K Campbell, B Lee (1991) “Name generators in surveys of personal networks.” Social Networks.
- M Resnick et al. (1997) “Protecting adolescents from harm: findings from the national longitudinal study on adolescent health.” JAMA.
- C Apicella, F Marlowe, J Fowler, N Christakis (2012) “Social networks and cooperation in hunter-gatherers,” Nature.
- B Ball, MEJ Newman (2013) “Friendship networks and social status.” Network Science.

Online Social Networks

S Acquaintances, often international. Business and personal.

f 2004: classmates, 2015: most people you know who are online.

in Mostly professional connections, some friends.

Virtual acquaintances, often interest-driven.

flickr Photography-interested real-world/virtual friends.

orange™ People you talk to on the phone, including customer service.

+ Close friends who exercise.

Sometimes personal, sometimes professional, sometimes both.

Online Social Networks

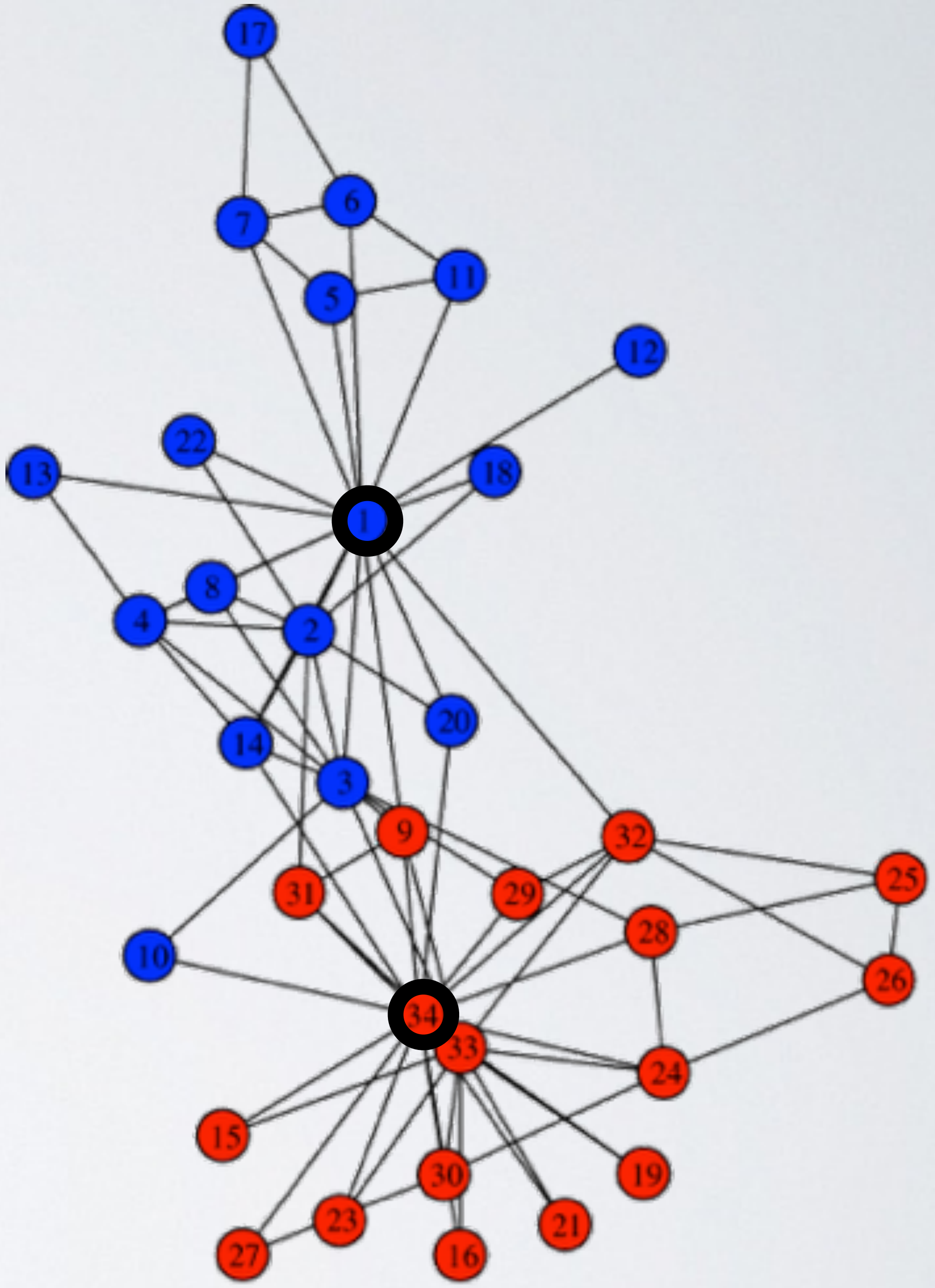
Some differences:

- Design aspects
- Personal vs. professional
- Strong vs. weak (Onnela et al. 2007)
- Virtual/real-world acquaintances (Jacobs et al. 2015)
- Single interest vs. diverse interest networks
- Co-tag friends vs. news feed friends vs. chat friends
- Phone calls vs. texts
- …

- JP Onnela et al. (2007) "Structure and tie strengths in mobile communication networks," PNAS.
- AZ Jacobs, SF Way, J Ugander, A Clauset (2015) "Assembling thefacebook: Using Heterogeneity to Understand Online Social Network Assembly," WebSci.

Zachary Karate Club

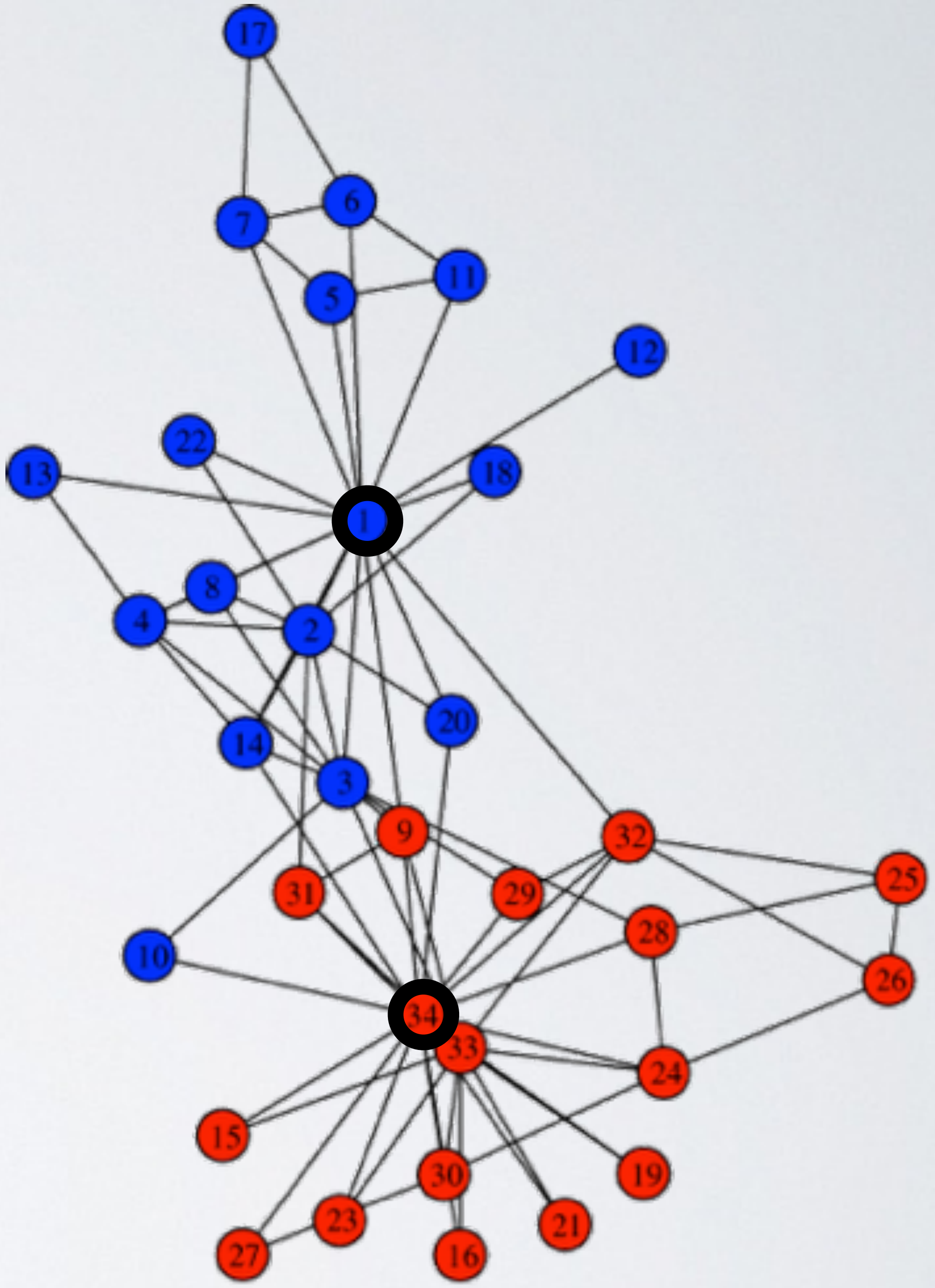
- Wayne Zachary, sociologist interested in group dynamics.
- Edges: interacted outside the club
- Studied a karate club for 3 years ('70–'72)
- Club formed factions around instructor (1) and Club President (34).
- Zachary was interested in if faction structure could be predicted.
- Method?



- W. W. Zachary (1977) "An information flow model for conflict and fission in small groups", J Anthro Research

Zachary Karate Club

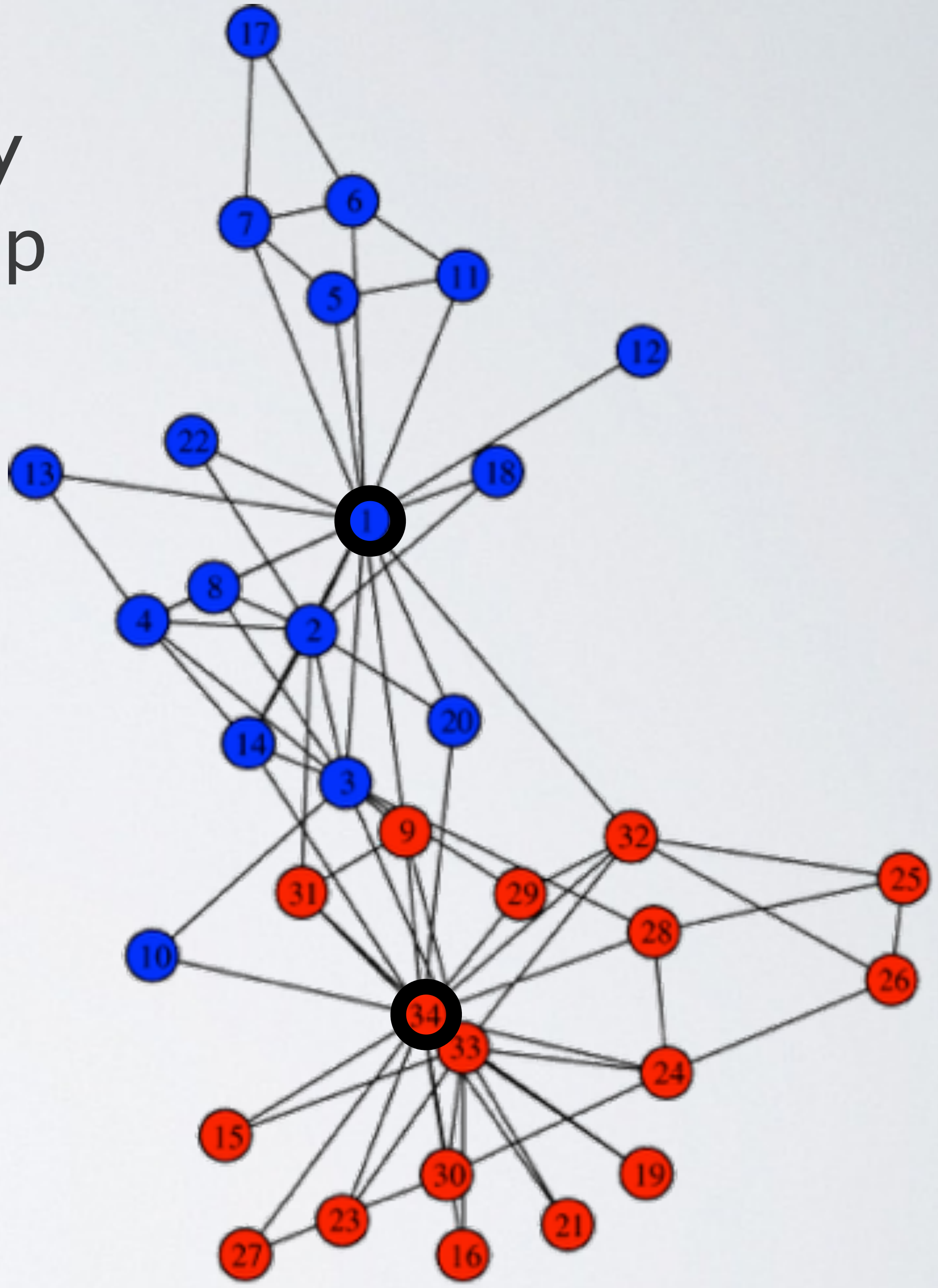
- Wayne Zachary, sociologist interested in group dynamics.
- Edges: interacted outside the club
- Studied a karate club for 3 years ('70-'72)
- Club formed factions around instructor (1) and Club President (34).
- Zachary was interested in if faction structure could be predicted.
- Method?
- **Network Flow!** Applied Ford-Fulkerson, found group split was predicted by min-cut.



- W. W. Zachary (1977) "An information flow model for conflict and fission in small groups", J Anthro Research

From Karate to Communities

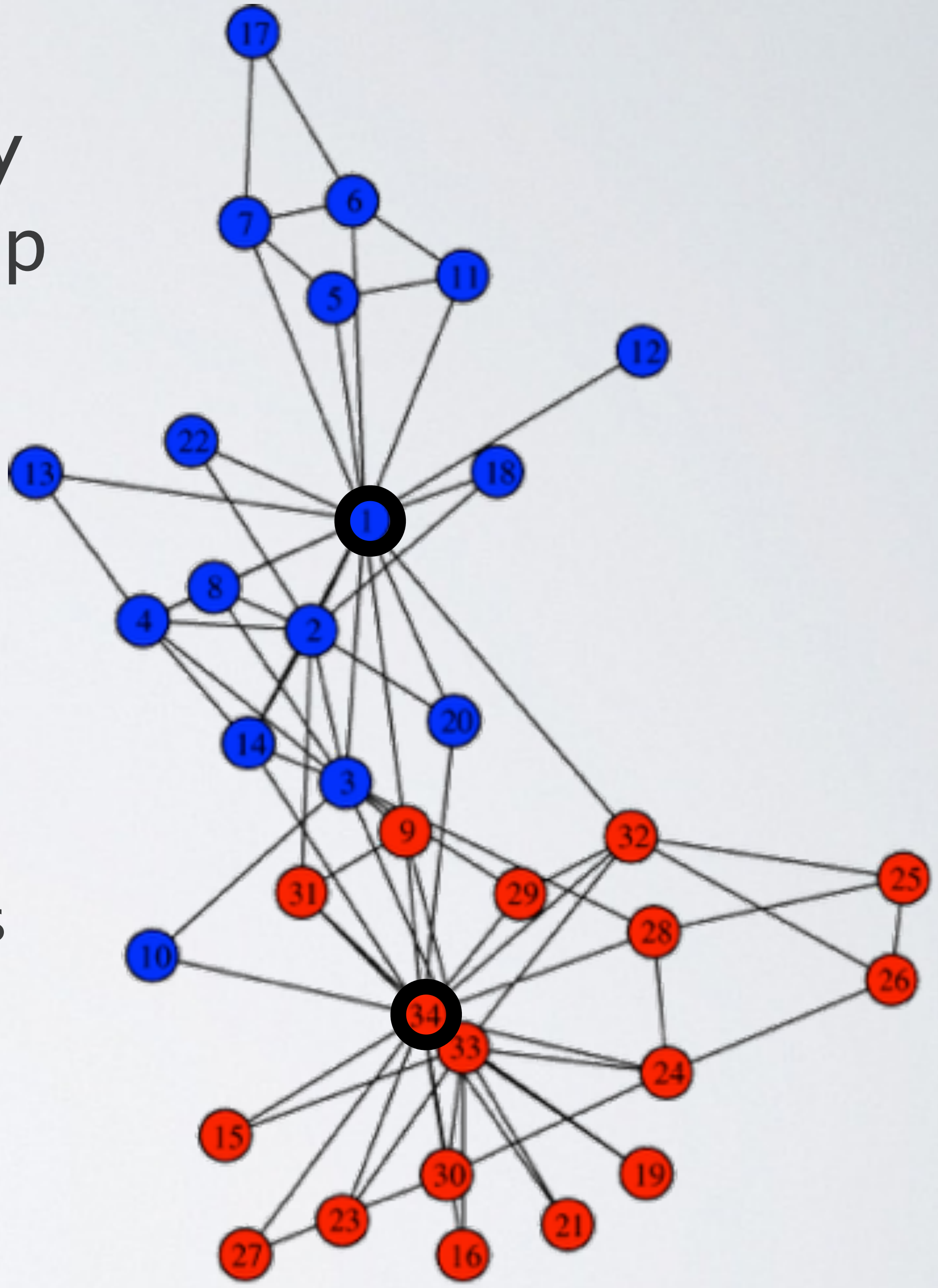
- **Zachary "objective function" for community detection**: does algorithm predict how a group fissions when led by two rival leaders?



- W. W. Zachary (1977) "An information flow model for conflict and fission in small groups", J Anthro Research

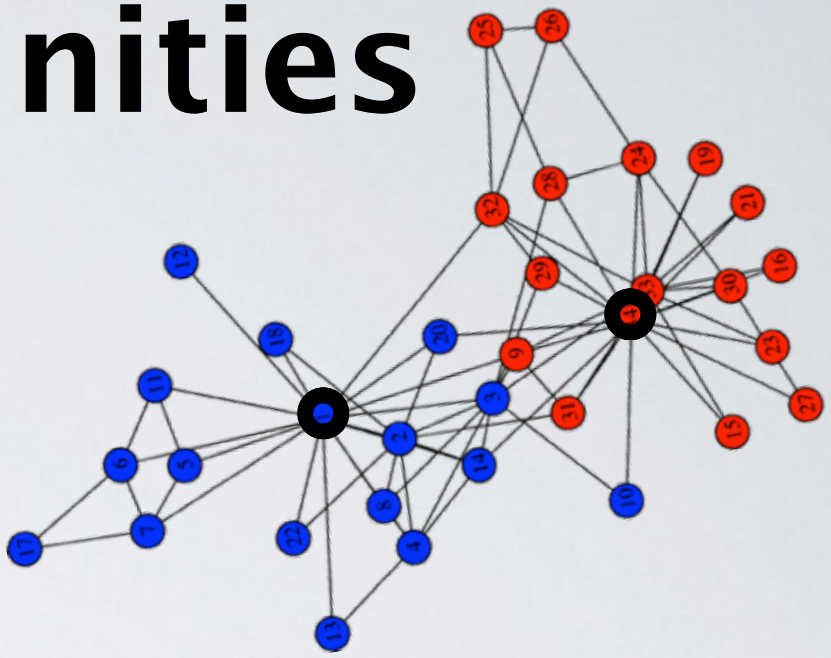
From Karate to Communities

- **Zachary "objective function" for community detection**: does algorithm predict how a group fissions when led by two rival leaders?
- Other objectives
  - Modularity maximization:
    - Has "resolution limit"
  - Conductance (normalized min-cut):
    - Produces balanced partitions; spectral guarantees

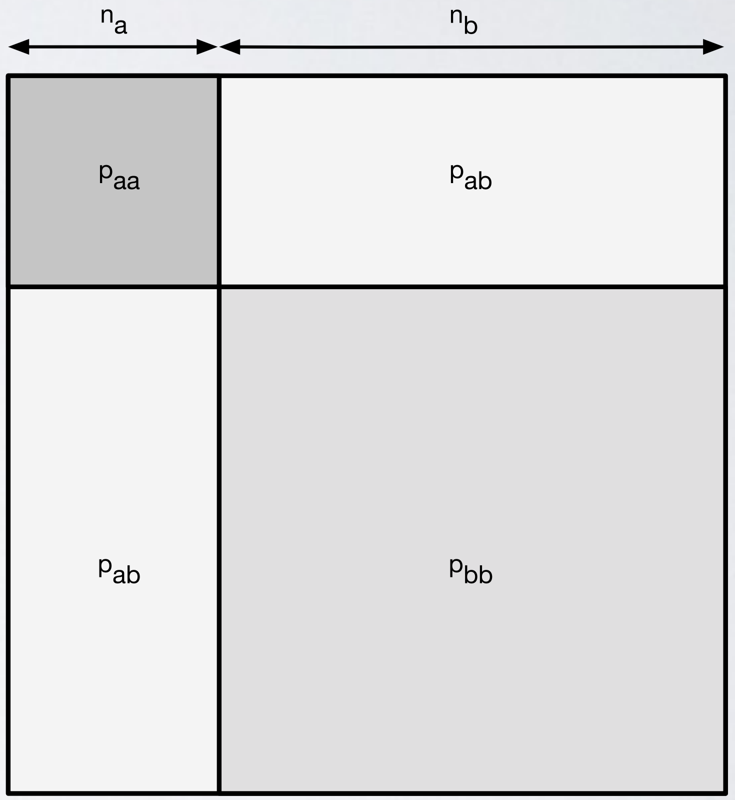


- MEJ Newman, M Girvan (2004) "Finding and evaluating community structure in networks," Physical Rev E.
- S Fortunato, M Barthelemy (2007) "Resolution limit in community detection," PNAS.
- J Shi, J Malik (2000) "Normalized cuts and image segmentation," IEEE Trans Pattern Analysis and Machine Intelligence.
- E Mossel, J Neeman, A Sly (2012) "Stochastic block models and reconstruction"

From Karate to Communities

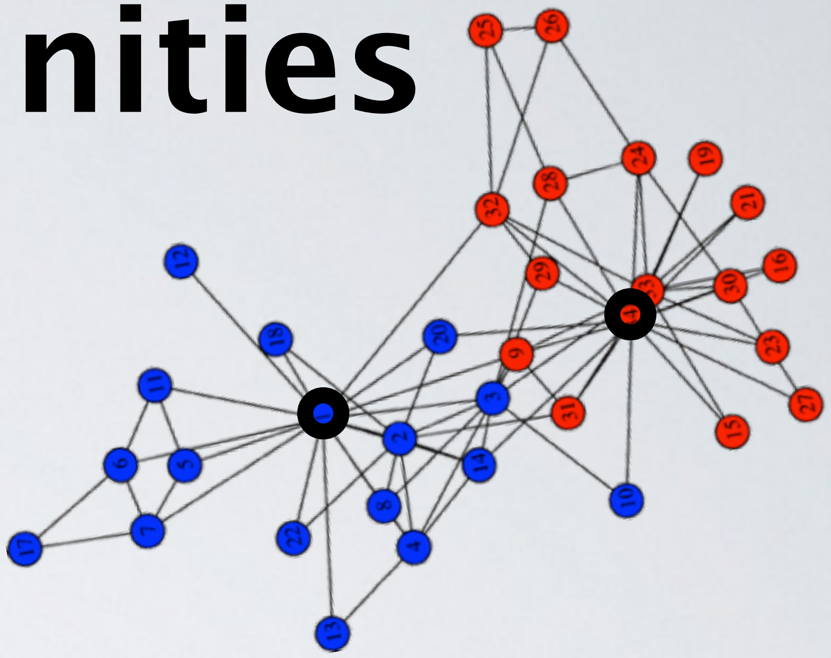


- **Zachary “objective function” for community detection**: does algorithm predict how a group fissions when led by two rival leaders?
- Other objectives
  - Modularity maximization:
    - Has “resolution limit”
  - Conductance (normalized min-cut):
    - Produces balanced partitions; spectral guarantees
  - Ability to recover Stochastic Block Model:
    - Stylized model in absence of ground truth data

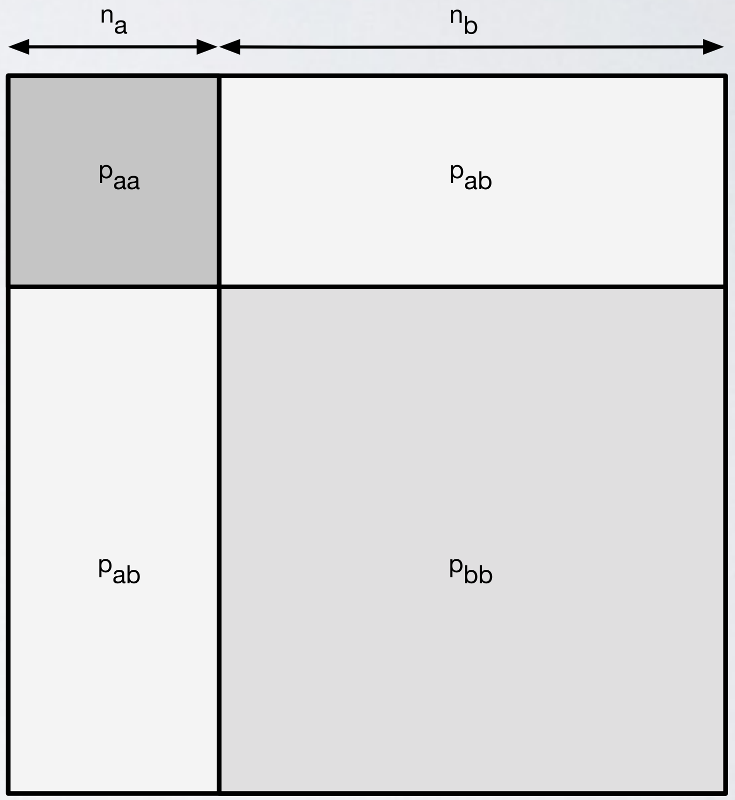


- MEJ Newman, M Girvan (2004) "Finding and evaluating community structure in networks,” Physical Rev E.
- S Fortunato, M Barthelemy (2007) "Resolution limit in community detection," PNAS.
- J Shi, J Malik (2000) “Normalized cuts and image segmentation,” IEEE Trans Pattern Analysis and Machine Intelligence.
- E Mossel, J Neeman, A Sly (2012) “Stochastic block models and reconstruction”

From Karate to Communities



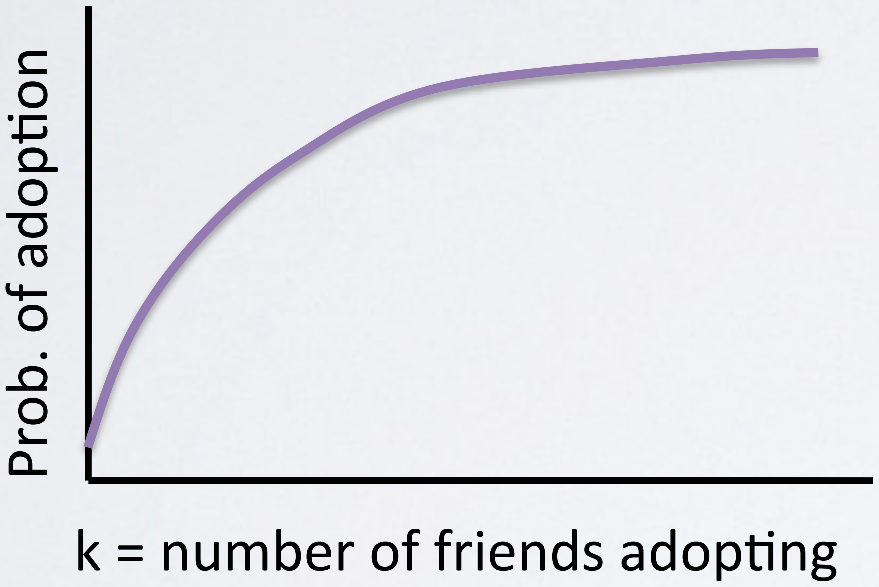
- **Zachary "objective function" for community detection**: does algorithm predict how a group fissions when led by two rival leaders?
- Other objectives
  - Modularity maximization:
    - Has "resolution limit"
  - Conductance (normalized min-cut):
    - Produces balanced partitions; spectral guarantees
  - Ability to recover Stochastic Block Model:
    - Stylized model in absence of ground truth data
  - Use in clustered network experiments?

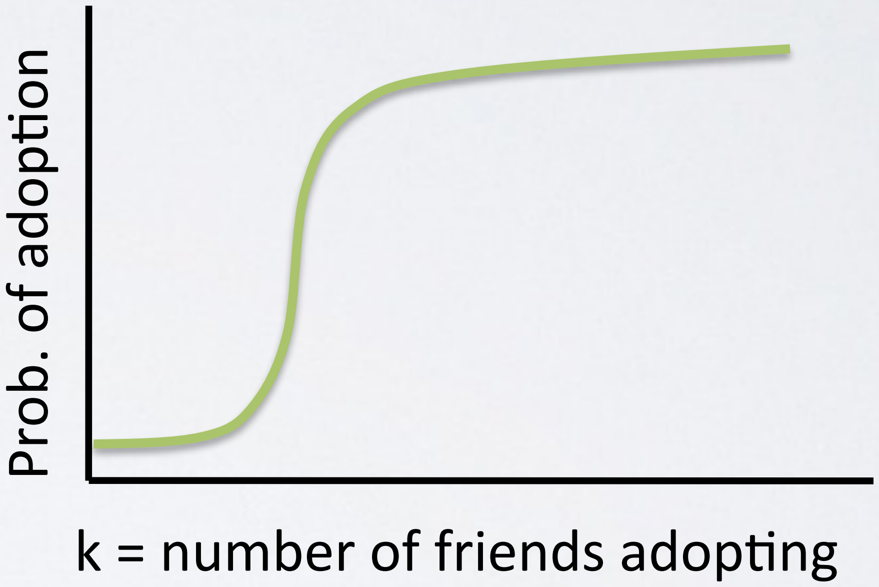


- MEJ Newman, M Girvan (2004) "Finding and evaluating community structure in networks," Physical Rev E.
- S Fortunato, M Barthelemy (2007) "Resolution limit in community detection," PNAS.
- J Shi, J Malik (2000) "Normalized cuts and image segmentation," IEEE Trans Pattern Analysis and Machine Intelligence.
- E Mossel, J Neeman, A Sly (2012) "Stochastic block models and reconstruction"

Influence, instrumented

- Prob. of adoption depends on the number of friends who have adopted (Bass 1969, Granovetter 1978)
- What is the shape? Diminishing returns? Critical mass?

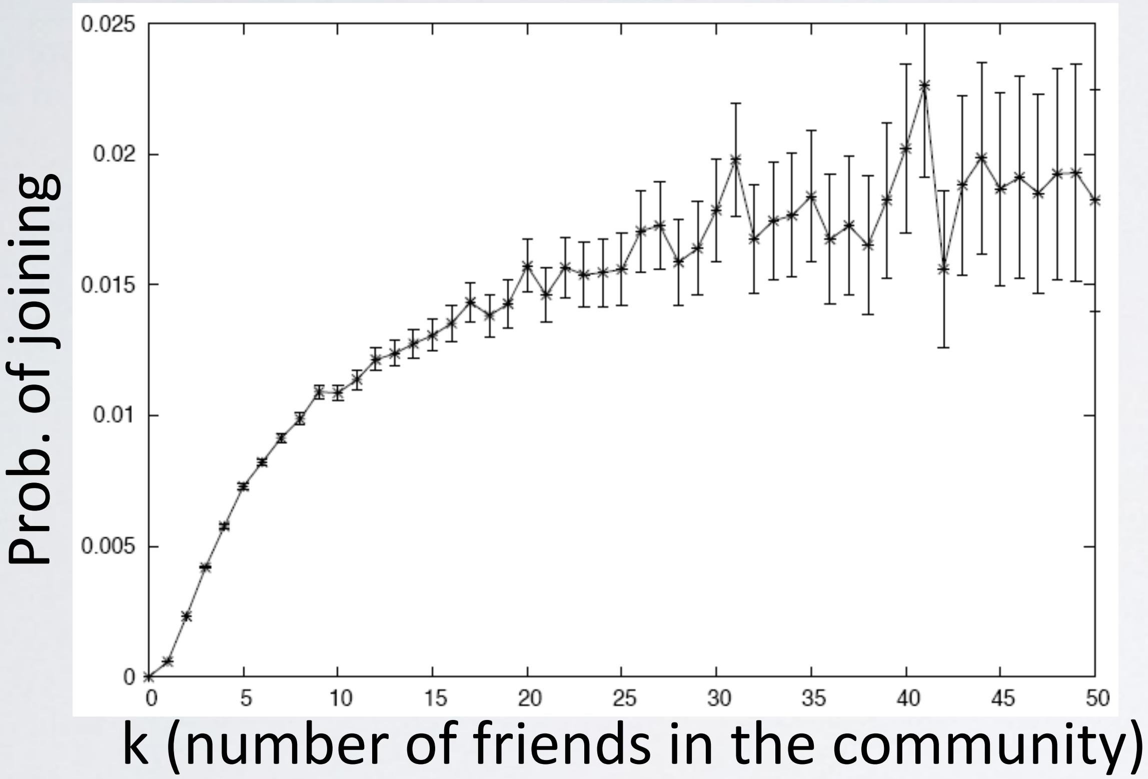




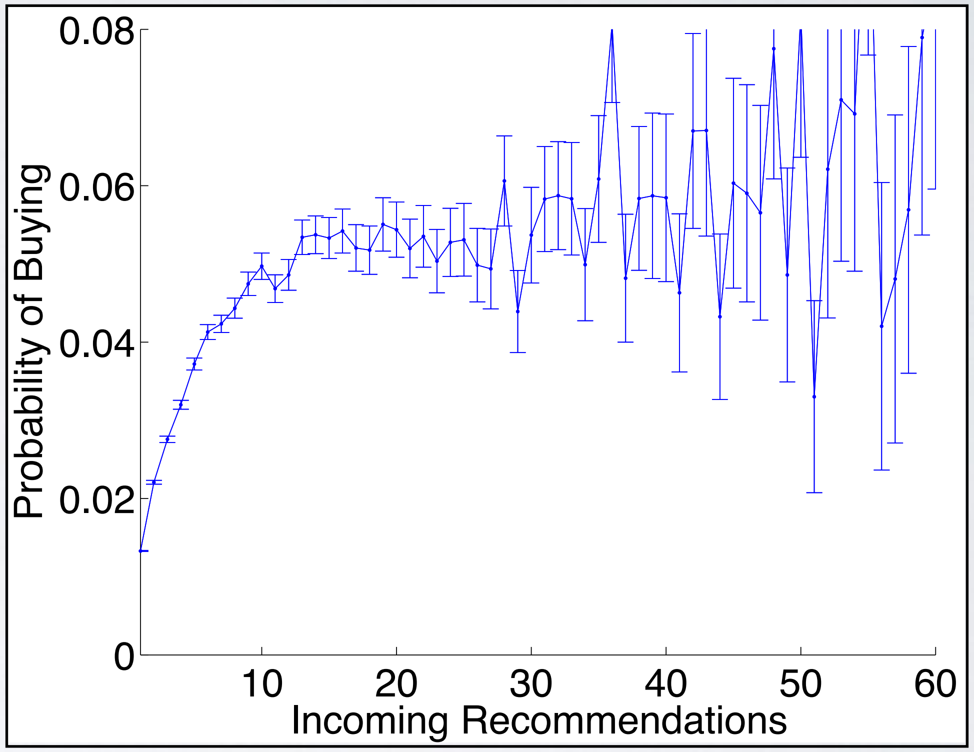
- F Bass (1969) "A new product growth for model consumer durables". Management Science.
- M Granovetter (1978) "Threshold models of collective action," American Journal Sociology.
- D Watts, P Dodds (2007) “Influentials, Networks, and Public Opinion Formation” Journal of Consumer Research.

Influence, instrumented

Backstrom et al. 2006: Probability of joining LiveJournal group



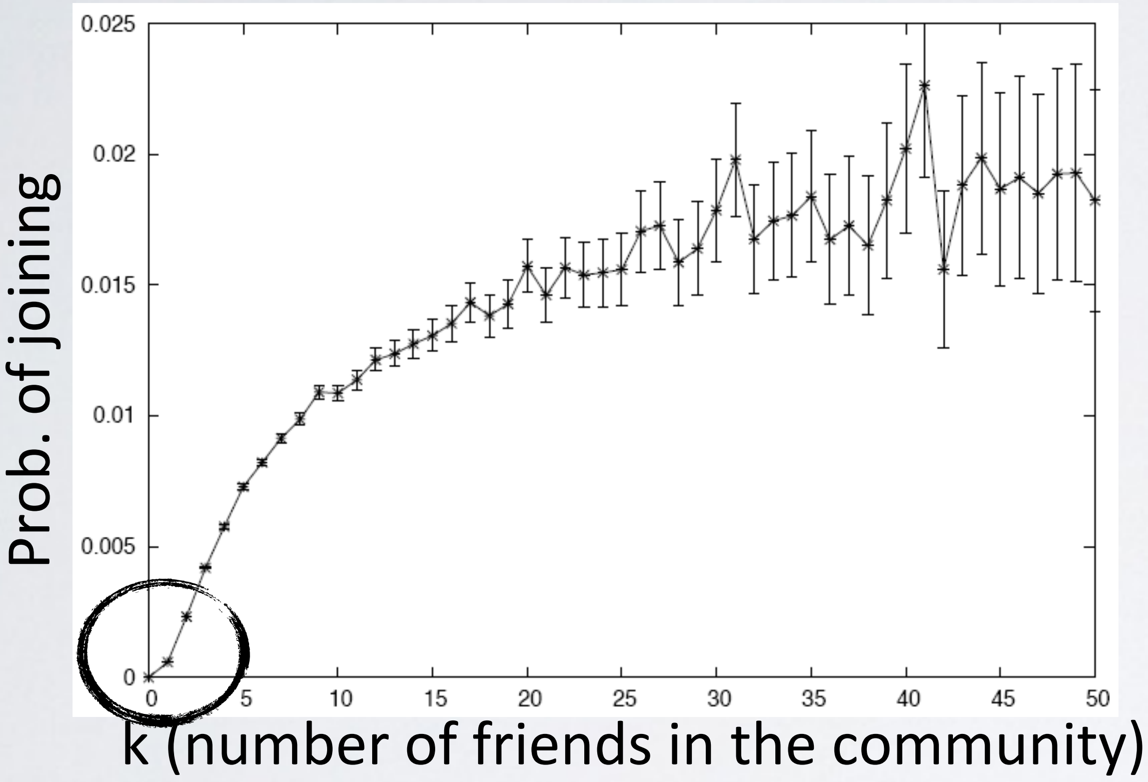
Leskovec et al. 2006: Probability of buying a DVD



- L Backstrom, D Huttenlocher, J Kleinberg, X Lan (2006) "Group formation in large social networks: membership, growth, and evolution," KDD.
- J Leskovec, LA Adamic, BA Huberman (2006) "The dynamics of viral marketing," EC.
- D Centola, V Eguiluz, M Macy (2007) "Cascade dynamics of complex propagation," Physica A.
- D Centola, M Macy (2007) "Complex contagions and the weakness of long ties" American Journal Sociology.

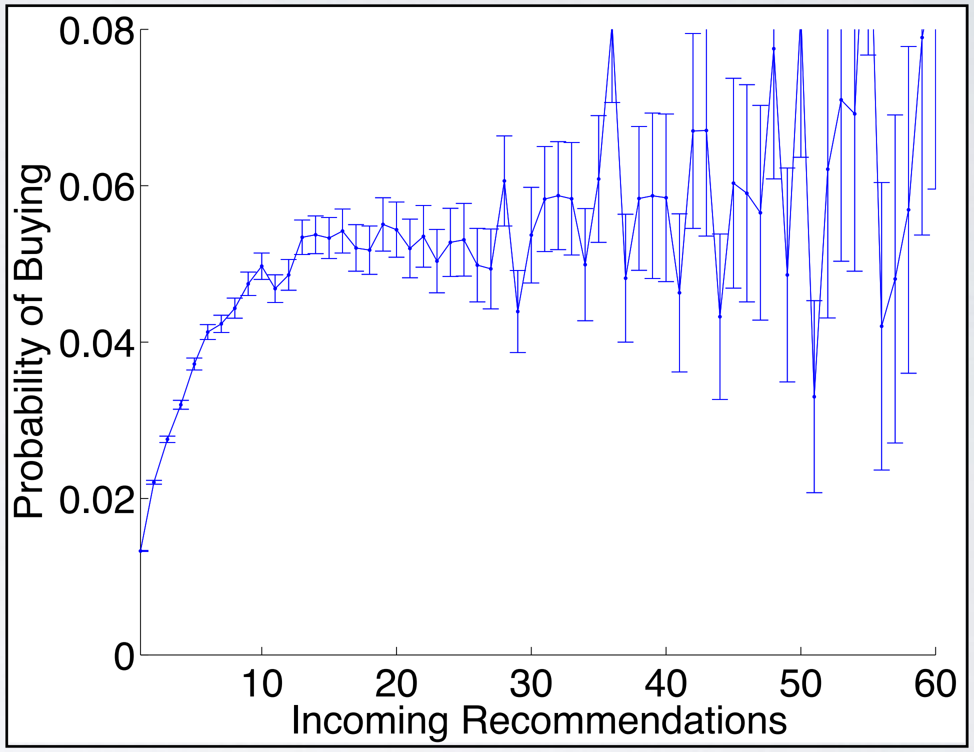
Influence, instrumented

Backstrom et al. 2006: Probability of joining LiveJournal group



Complex contagion?

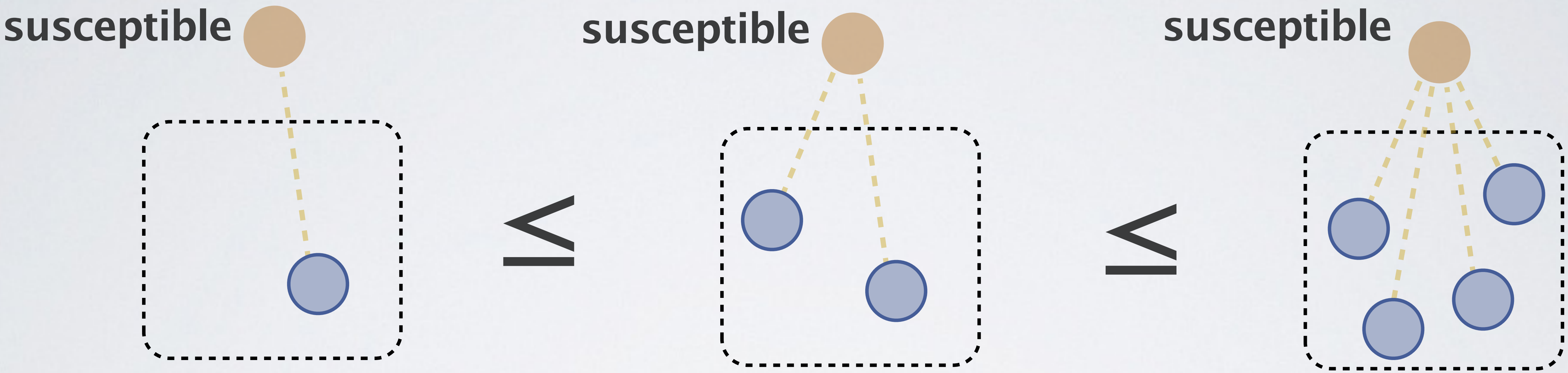
Leskovec et al. 2006: Probability of buying a DVD



- L Backstrom, D Huttenlocher, J Kleinberg, X Lan (2006) "Group formation in large social networks: membership, growth, and evolution," KDD.
- J Leskovec, LA Adamic, BA Huberman (2006) "The dynamics of viral marketing," EC.
- D Centola, V Eguiluz, M Macy (2007) "Cascade dynamics of complex propagation," Physica A.
- D Centola, M Macy (2007) "Complex contagions and the weakness of long ties" American Journal Sociology.

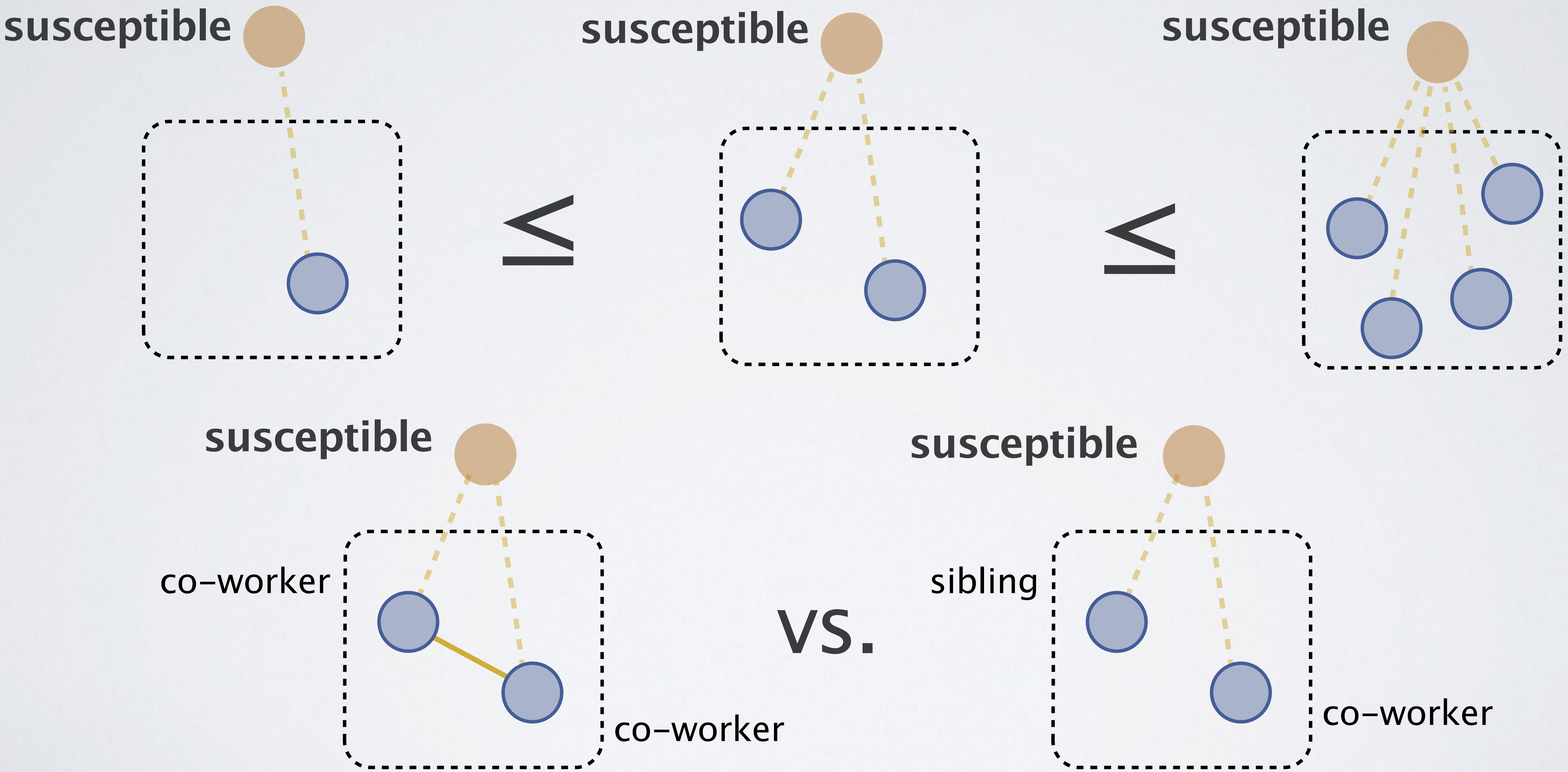
Influence and graph structure

- Adoption as a simple function of ‘contact neighborhood’ size:



Influence and graph structure

- Adoption as a simple function of ‘contact neighborhood’ size:



susceptible ≤ susceptible ≤ susceptible

susceptible

co–worker

co–worker

vs.

susceptible

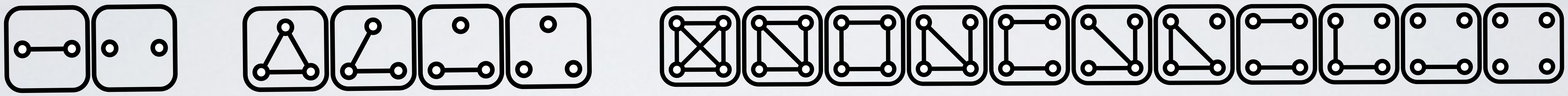
sibling

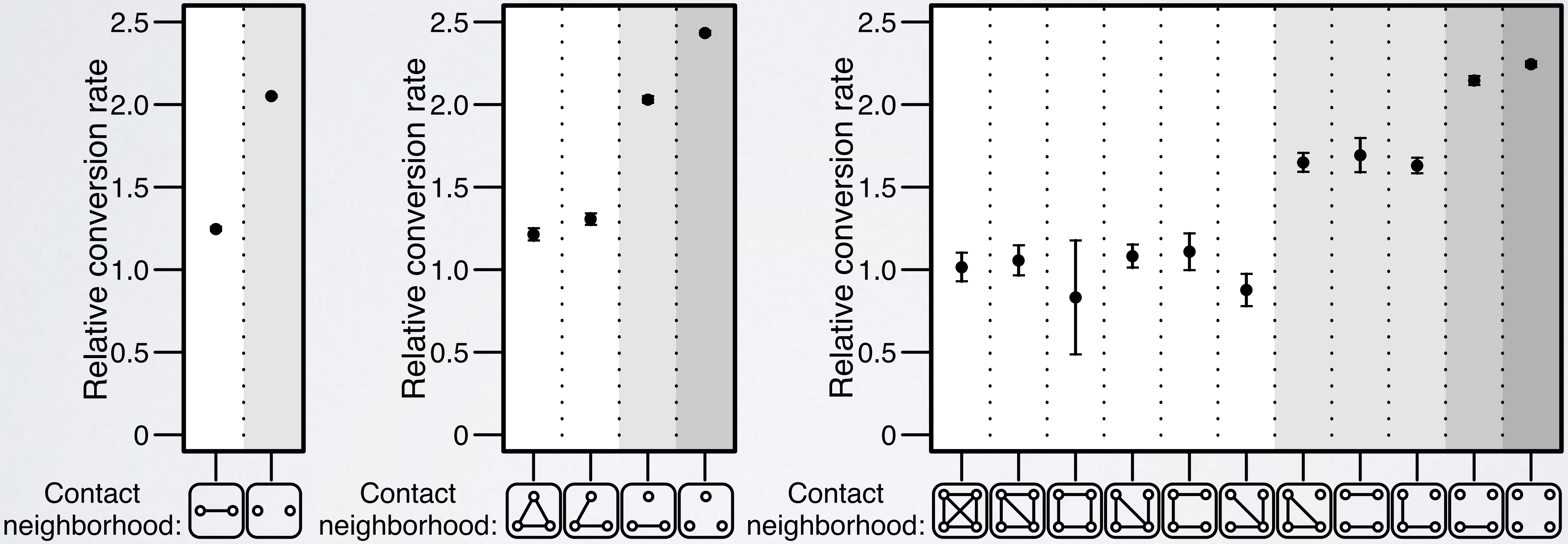
co–worker

Structural diversity

Conversion rate on invitations to Facebook as a function of graph, “f(G)”?

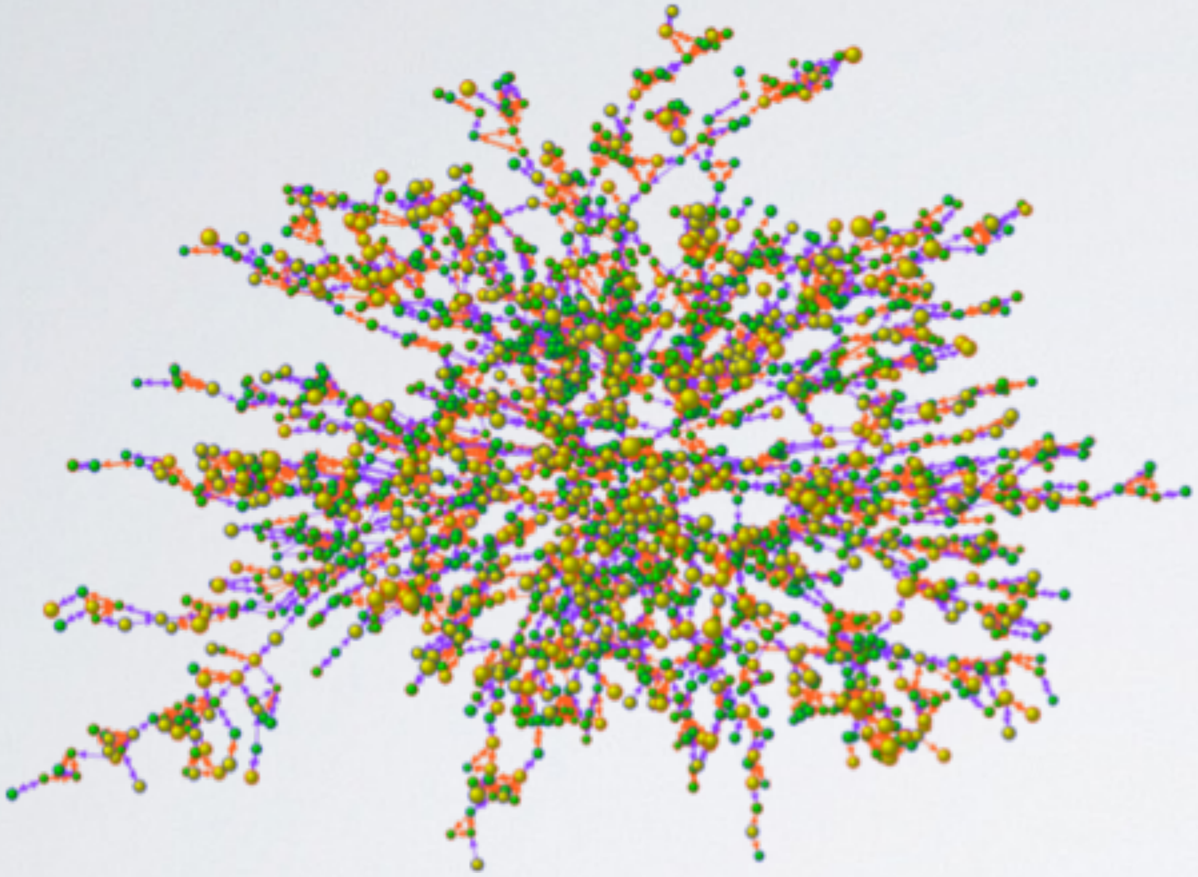
size 2: size 3: size 4:

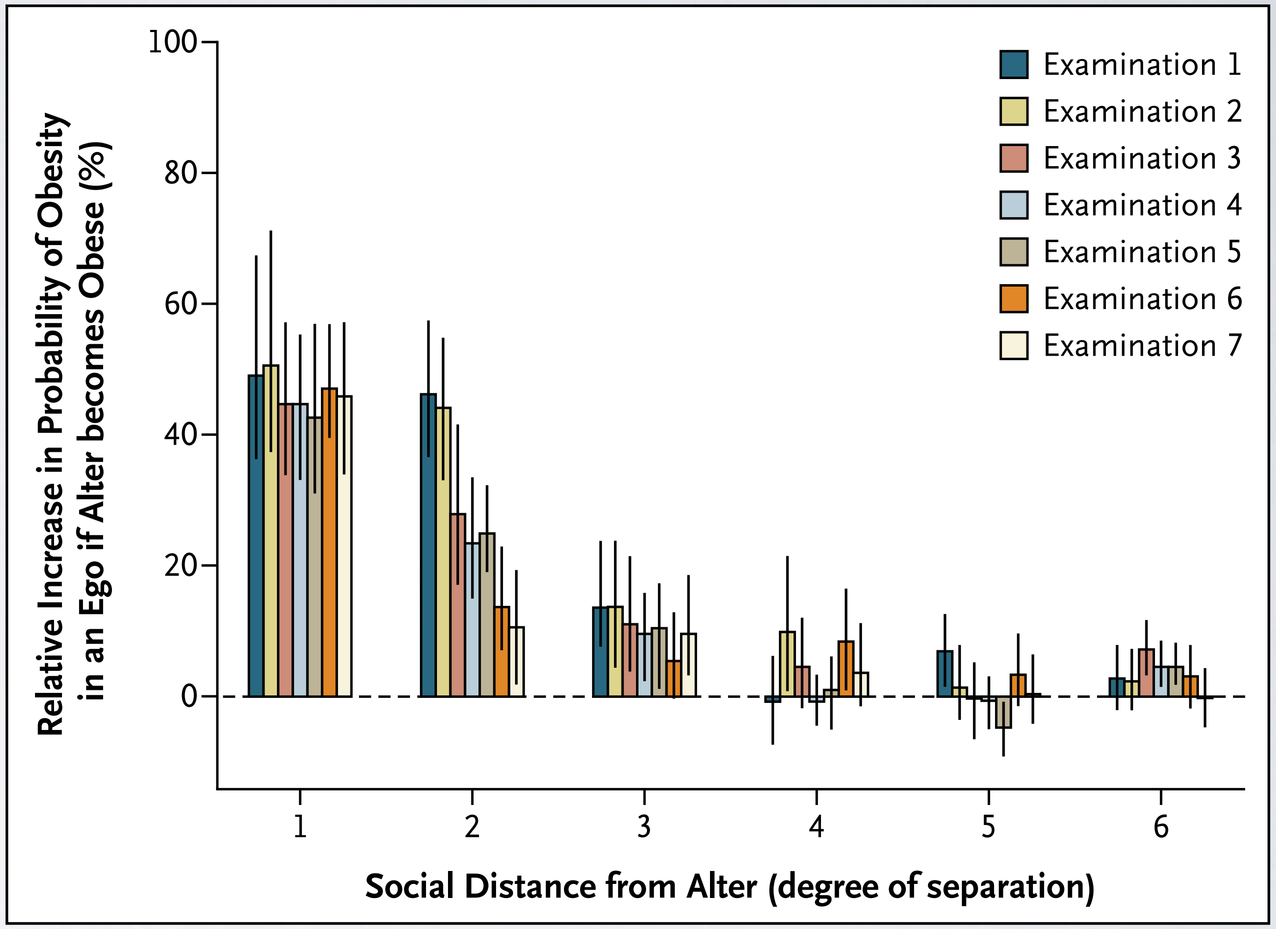




- J Ugander, L Backstrom, C Marlow, J Kleinberg (2012) “Structural diversity in social contagion,” PNAS.

Is obesity contagious?

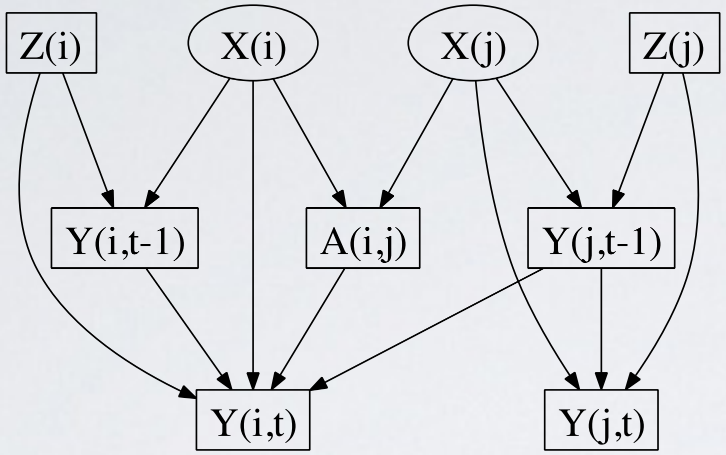



100
80
60
40
20
0
Relative Increase in Probability of Obesity in an Ego if Alter becomes Obese (%)
Examination 1
Examination 2
Examination 3
Examination 4
Examination 5
Examination 6
Examination 7
1
2
3
4
5
6
Social Distance from Alter (degree of separation)

“comparing the conditional probability of obesity in the observed network with the probability of obesity in identical networks (with topology preserved) in which the same number of obese persons is randomly distributed”

- N Christakis, J Fowler (2007) "The Spread of Obesity in a Large Social Network over 32 Years," New England J of Medicine.
- C Shalizi, A Thomas (2011) "Homophily and contagion are generically confounded in observational social network studies," Sociological Methods & Research.

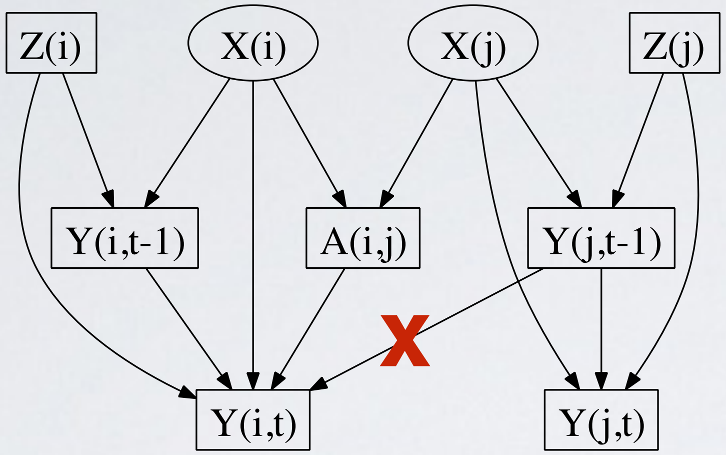
Knock-out experiments



Symbol	Meaning
i, j	Individuals
Z	Observed Traits
X	Latent Traits
Y	Observed Outcomes

- N Christakis, J Fowler (2007) "The Spread of Obesity in a Large Social Network over 32 Years," New England J of Medicine.
- C Shalizi, A Thomas (2011) "Homophily and contagion are generically confounded in observational social network studies," Sociological Methods & Research.
- E Bakshy et al. (2012) “The Role of Social Networks in Information Diffusion,” WWW.

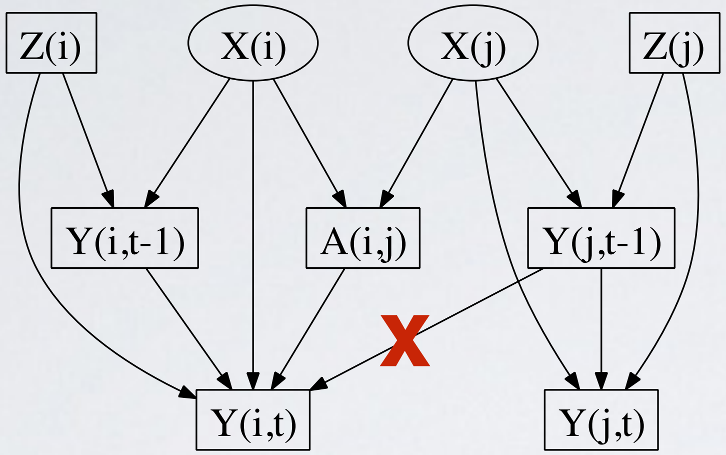
Knock-out experiments



Symbol	Meaning
i, j	Individuals
Z	Observed Traits
X	Latent Traits
Y	Observed Outcomes

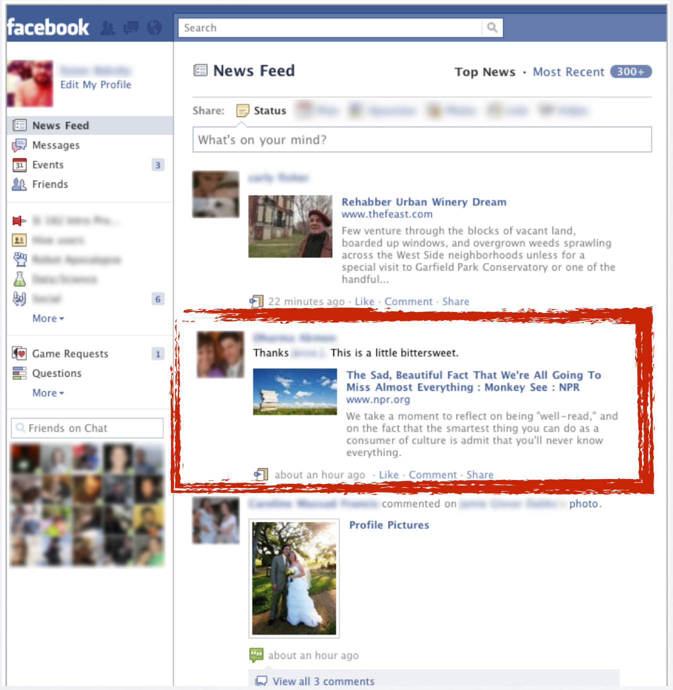
- N Christakis, J Fowler (2007) "The Spread of Obesity in a Large Social Network over 32 Years," New England J of Medicine.
- C Shalizi, A Thomas (2011) "Homophily and contagion are generically confounded in observational social network studies," Sociological Methods & Research.
- E Bakshy et al. (2012) “The Role of Social Networks in Information Diffusion,” WWW.

Knock-out experiments



Symbol	Meaning
i, j	Individuals
Z	Observed Traits
X	Latent Traits
Y	Observed Outcomes

“Feed Condition”



News Feed

[redacted]

Rehabber Urban Winery Dream
www.thefeast.com
Few venture through the blocks of vacant land, boarded up windows, and overgrown weeds sprawling across the West Side neighborhoods unless for a special visit to Garfield Park Conservatory or one of the handful...

22 minutes ago · Like · Comment · Share

[redacted]
Thanks [redacted] This is a little bittersweet.

The Sad, Beautiful Fact That We're All Going To Miss Almost Everything : Monkey See : NPR
www.npr.org
We take a moment to reflect on being "well-read," and on the fact that the smartest thing you can do as a consumer of culture is admit that you'll never know everything.

about an hour ago · Like · Comment · Share

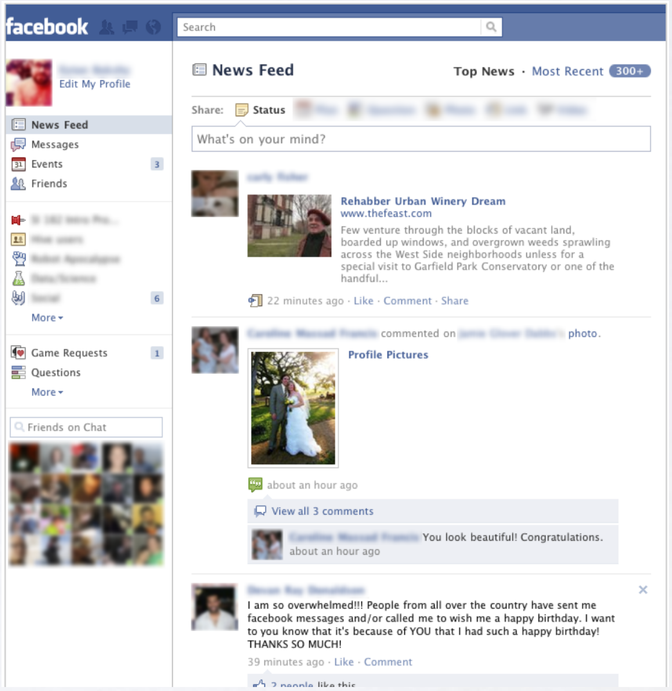
[redacted] commented on [redacted] photo.

Profile Pictures

about an hour ago

View all 3 comments

“No Feed Condition”



News Feed

[redacted]

Rehabber Urban Winery Dream
www.thefeast.com
Few venture through the blocks of vacant land, boarded up windows, and overgrown weeds sprawling across the West Side neighborhoods unless for a special visit to Garfield Park Conservatory or one of the handful...

22 minutes ago · Like · Comment · Share

[redacted] commented on [redacted] photo.

Profile Pictures

about an hour ago

View all 3 comments

[redacted] You look beautiful! Congratulations.
about an hour ago

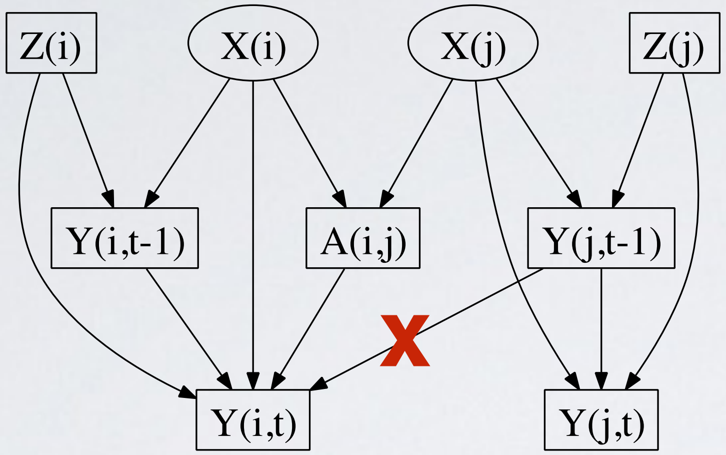
[redacted]
I am so overwhelmed!!! People from all over the country have sent me facebook messages and/or called me to wish me a happy birthday. I want to you know that it's because of YOU that I had such a happy birthday! THANKS SO MUCH!

39 minutes ago · Like · Comment

2 people like this

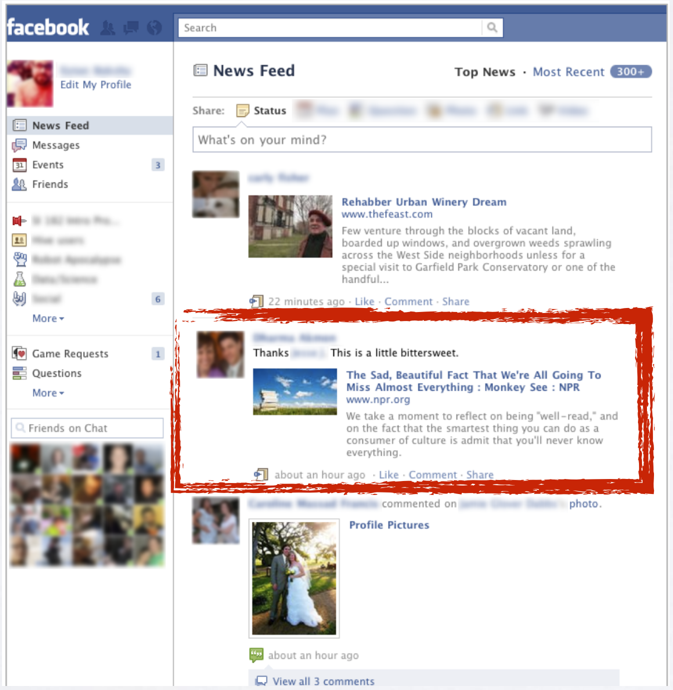
- N Christakis, J Fowler (2007) "The Spread of Obesity in a Large Social Network over 32 Years," New England J of Medicine.
- C Shalizi, A Thomas (2011) "Homophily and contagion are generically confounded in observational social network studies," Sociological Methods & Research.
- E Bakshy et al. (2012) “The Role of Social Networks in Information Diffusion,” WWW.

Knock-out experiments



Symbol	Meaning
i, j	Individuals
Z	Observed Traits
X	Latent Traits
Y	Observed Outcomes

“Feed Condition”

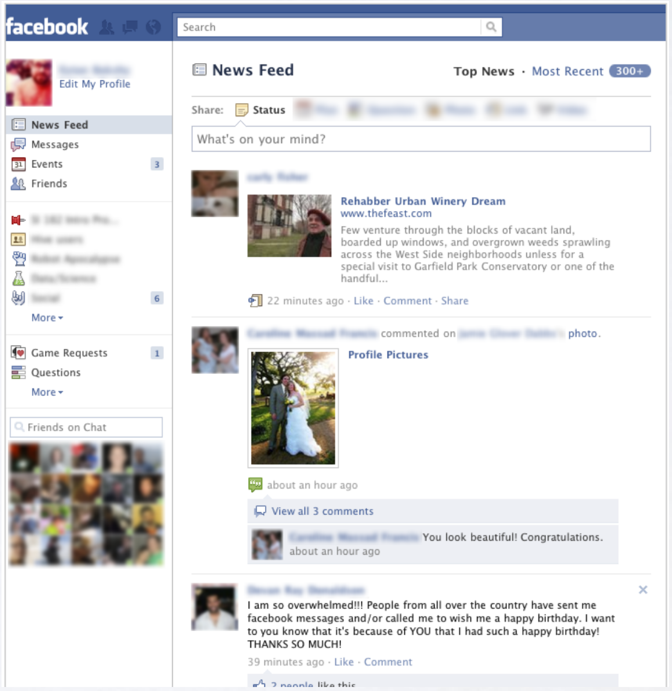


Rehabber Urban Winery Dream
www.thefeast.com
Few venture through the blocks of vacant land, boarded up windows, and overgrown weeds sprawling across the West Side neighborhoods unless for a special visit to Garfield Park Conservatory or one of the handful...
22 minutes ago · Like · Comment · Share

Thanks This is a little bittersweet.
The Sad, Beautiful Fact That We're All Going To Miss Almost Everything : Monkey See : NPR
www.npr.org
We take a moment to reflect on being "well-read," and on the fact that the smartest thing you can do as a consumer of culture is admit that you'll never know everything.
about an hour ago · Like · Comment · Share

commented on photo.
Profile Pictures
about an hour ago
View all 3 comments

“No Feed Condition”



Rehabber Urban Winery Dream
www.thefeast.com
Few venture through the blocks of vacant land, boarded up windows, and overgrown weeds sprawling across the West Side neighborhoods unless for a special visit to Garfield Park Conservatory or one of the handful...
22 minutes ago · Like · Comment · Share

commented on photo.
Profile Pictures
about an hour ago
View all 3 comments
You look beautiful! Congratulations.
about an hour ago

I am so overwhelmed!!! People from all over the country have sent me facebook messages and/or called me to wish me a happy birthday. I want to you know that it's because of YOU that I had such a happy birthday! THANKS SO MUCH!
39 minutes ago · Like · Comment
2 people like this

Feed condition:

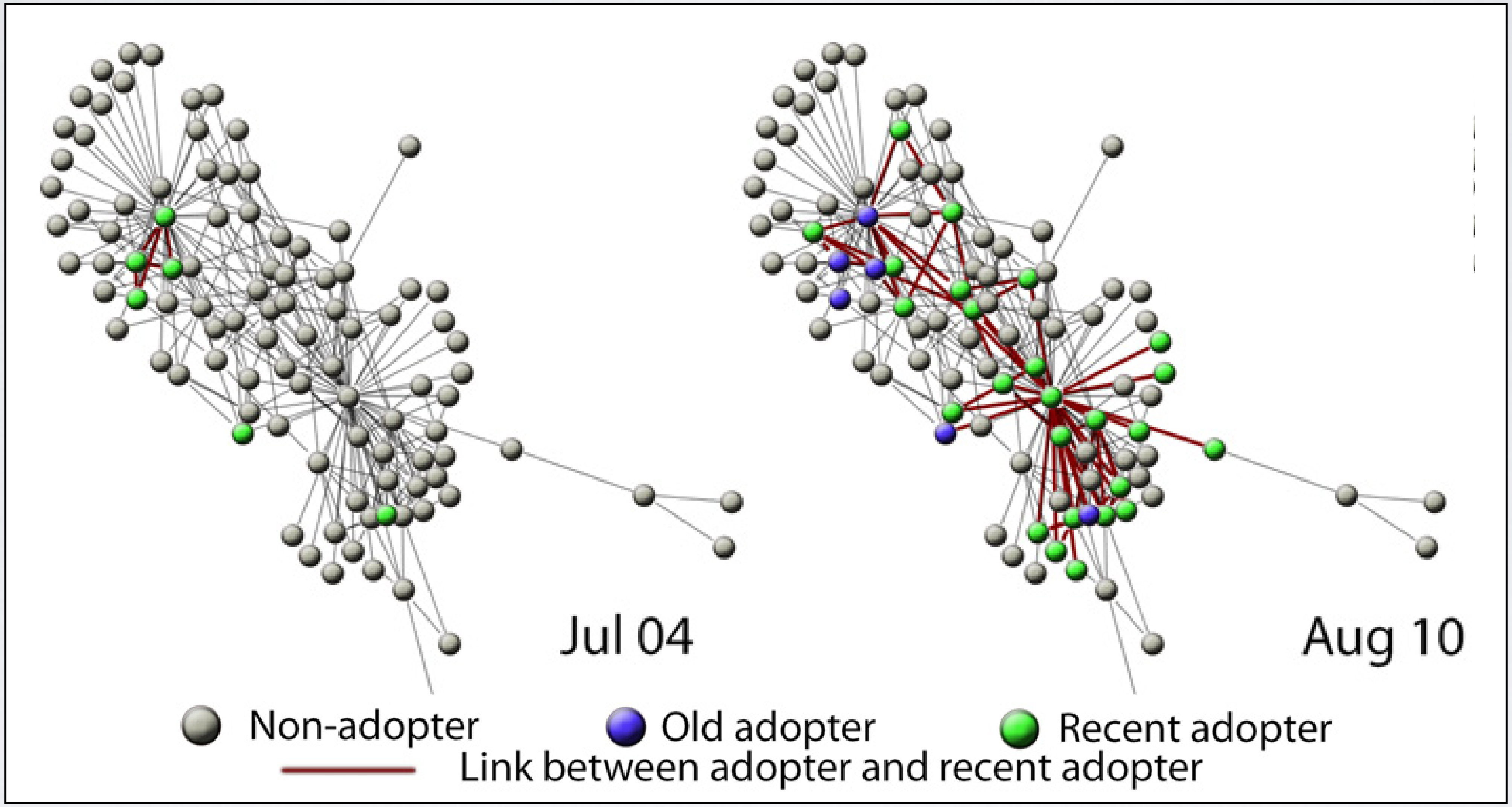
7.37x

more likely to share.

- N Christakis, J Fowler (2007) "The Spread of Obesity in a Large Social Network over 32 Years," New England J of Medicine.
- C Shalizi, A Thomas (2011) "Homophily and contagion are generically confounded in observational social network studies," Sociological Methods & Research.
- E Bakshy et al. (2012) “The Role of Social Networks in Information Diffusion,” WWW.

Do we need experiments?

- (Aral et al. 2009): Yahoo! Go service, 2007, n=27.4 million.

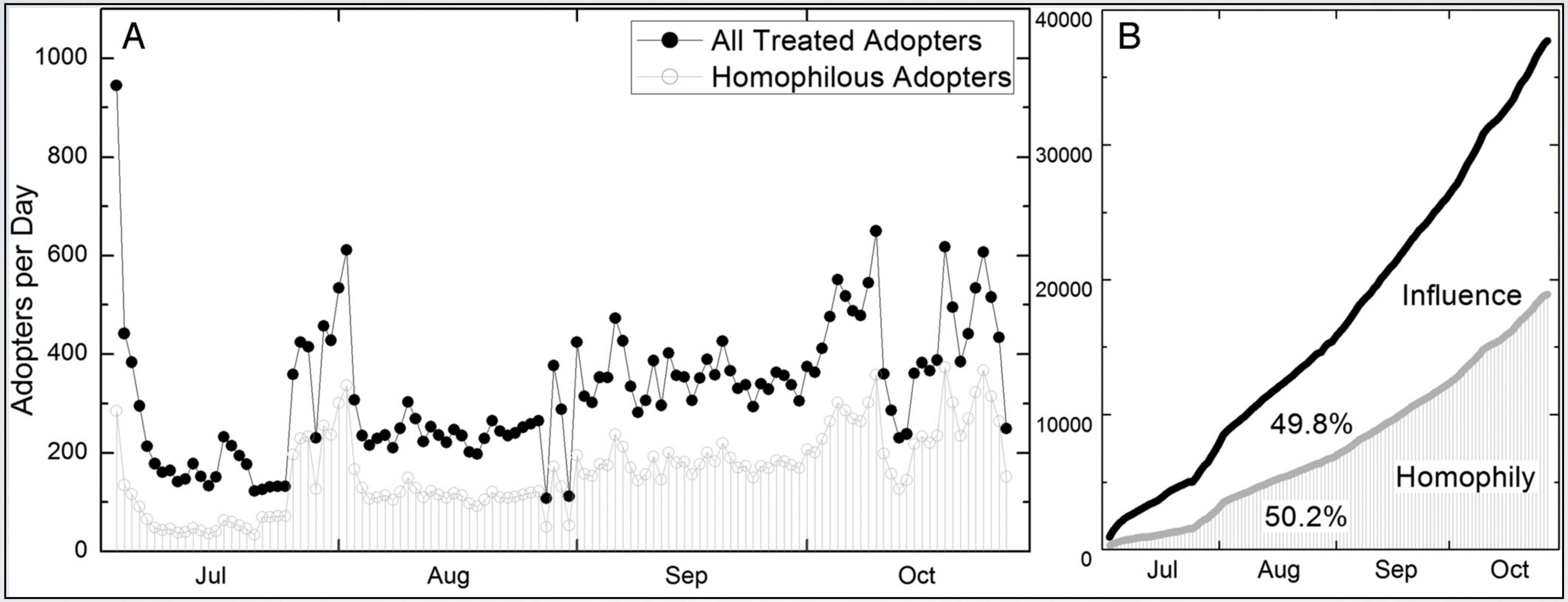


Is this social influence?

- P Rosenbaum, D Rubin (1983) "The central role of the propensity score in observational studies for causal effects," Biometrika.
- D Rubin (2006) “Matched sampling for causal effects”
- S Aral, L Muchnik, A Sundararajan (2009) "Distinguishing influence-based contagion from homophily-driven diffusion in dynami networks," PNAS.

Do we need experiments?

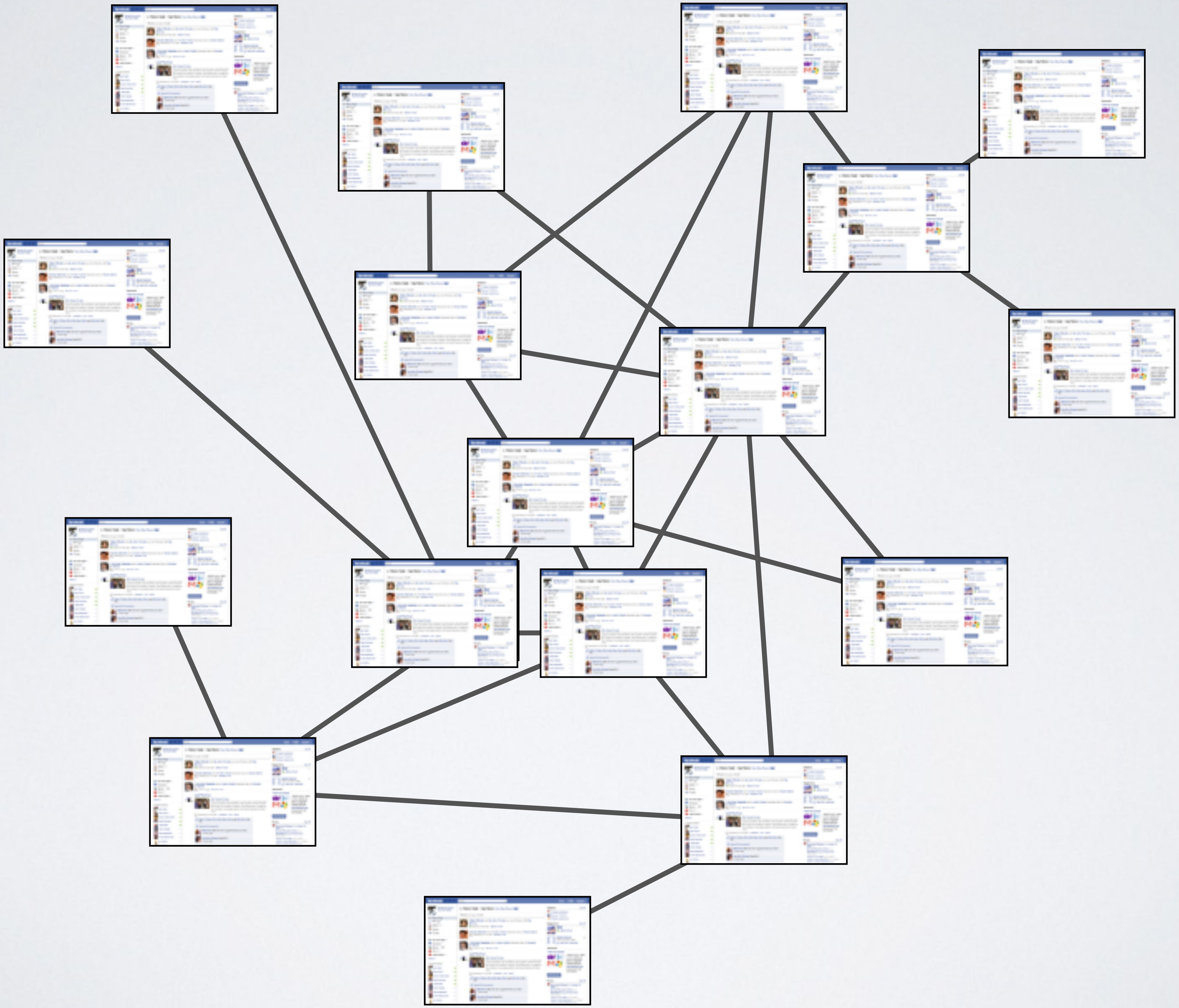
- (Aral et al. 2009): Yahoo! Go service, 2007, n=27.4 million.



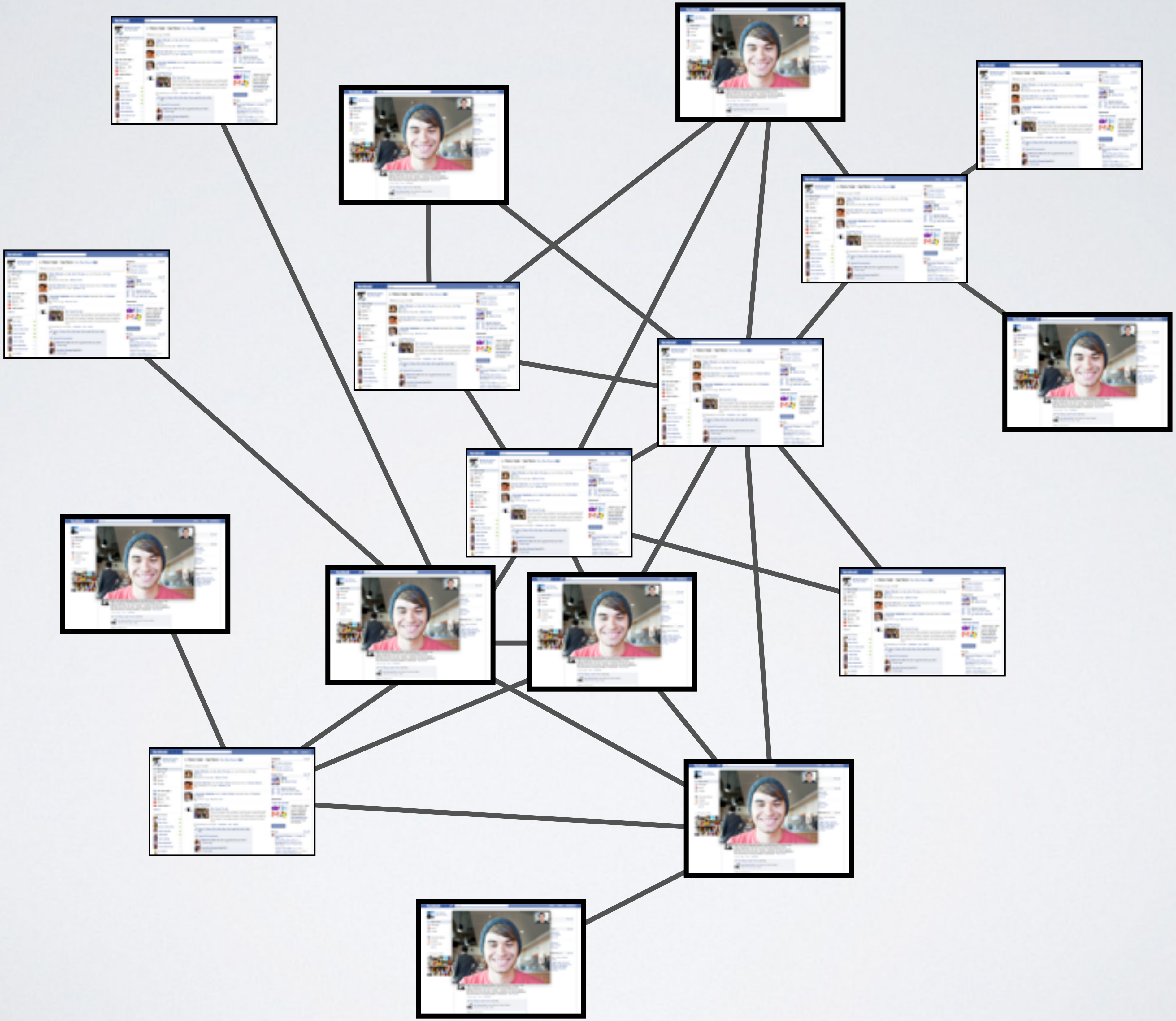
Is this social influence? ~50%

- P Rosenbaum, D Rubin (1983) "The central role of the propensity score in observational studies for causal effects," Biometrika.
- D Rubin (2006) “Matched sampling for causal effects”
- S Aral, L Muchnik, A Sundararajan (2009) "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," PNAS.

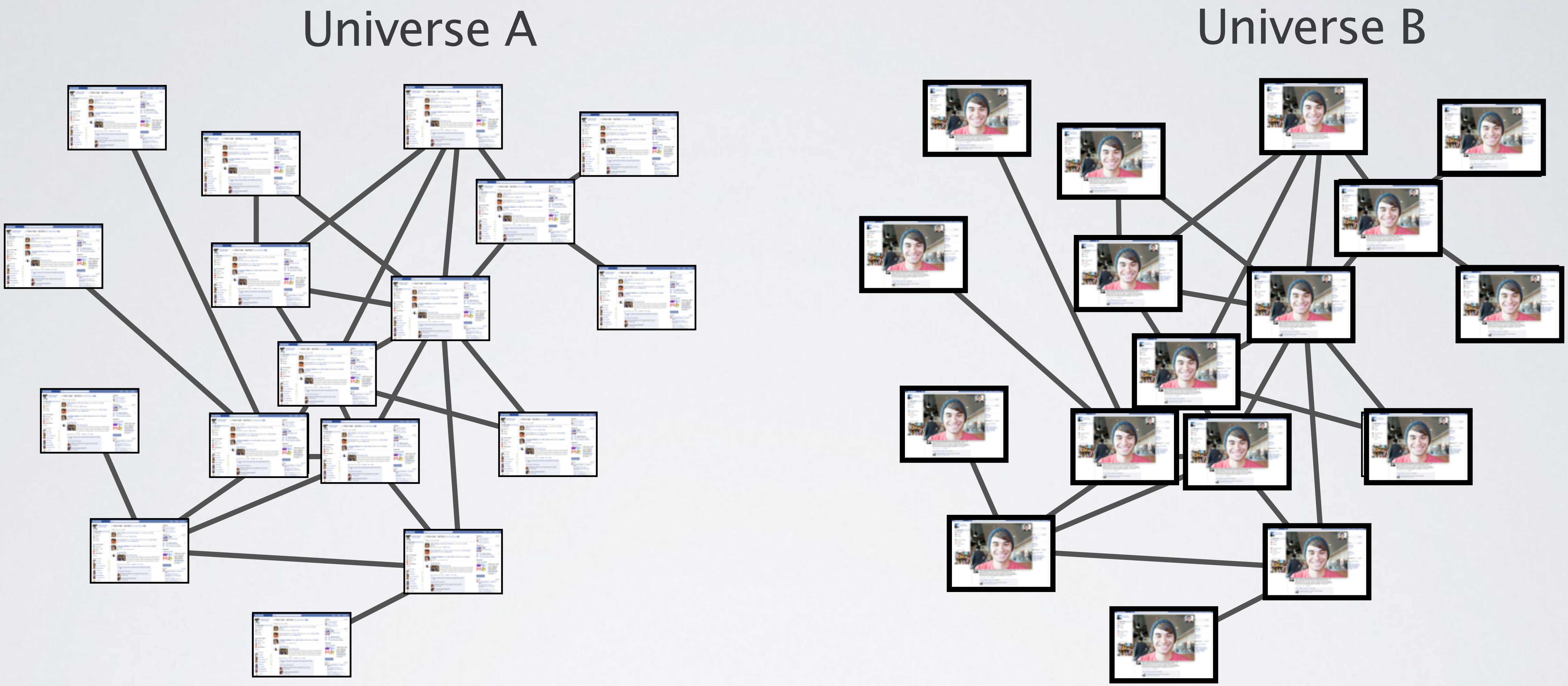
A/B testing under network effects



A/B testing under network effects



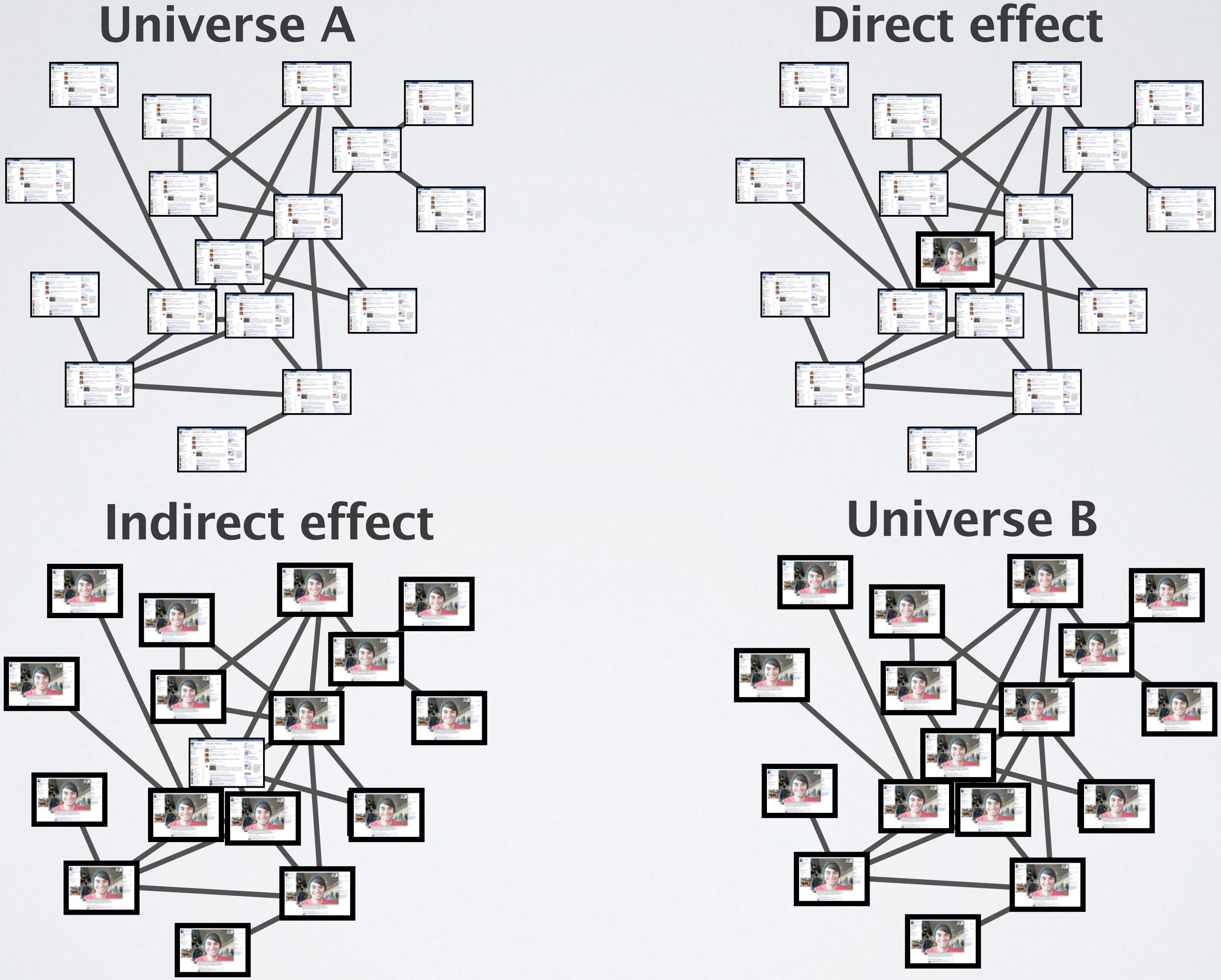
Causal inference & network effects



Fundamental problem: want to compare (average treatment effect, ATE), but can’t observe network in both states at once.

- J Ugander, B Karrer, L Backstrom, J Kleinberg (2013) "Graph Cluster Randomization: Network Exposure to Multiple Universes," KDD.
- D Eckles, B Karrer, J Ugander (2014) "Design and analysis of experiments in networks: Reducing bias from interference," arXiv.
- S Athey, D Eckles, G Imbens (2015) "Exact P-values for Network Interference," arXiv.

Direct vs. indirect effects



- P Aronow, C Samii (2013) "Estimating average causal effects under interference between units," arXiv.
- C Manski (2013) "Identification of treatment response with social interactions," The Econometrics Journal.

Learning outcomes

1. Students should develop a familiarity with relevant structural properties of empirical social networks, and how different graph models capture or don't capture these properties.

2. Students should be able to weigh advantages and disadvantages of different observational and experimental study designs that examine/test mechanisms in social systems.

3. Students should be able to employ structural measures for diverse ranking/predicting problems on graphs.

4. Students should be able to critically read research papers in the field to identify strengths and potential weaknesses, and to be able to design tests of potential weaknesses.