The Paradoxes of Social Data:

How Heterogeneity Distorts Information in Networks

Kristina Lerman

USC Information Sciences Institute

http://www.isi.edu/~lerman

Local vs Global

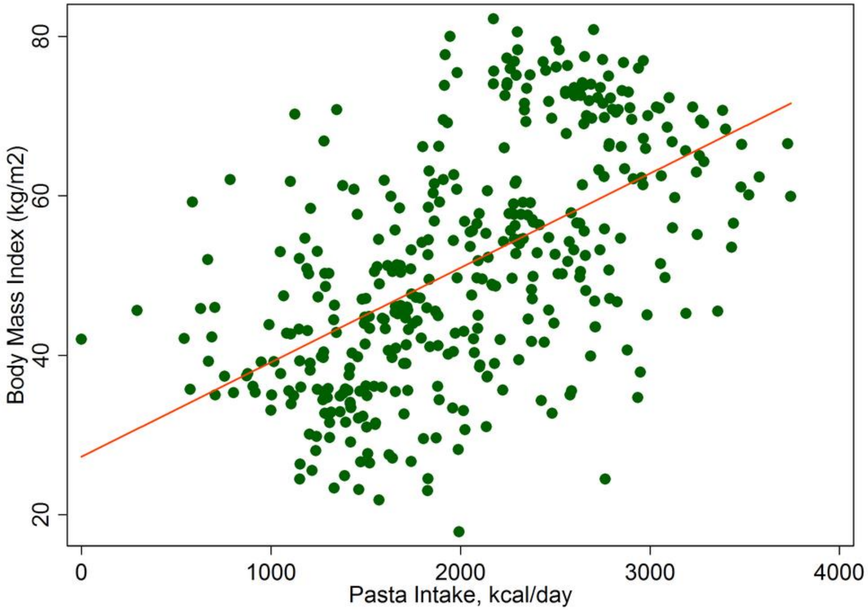
The local and global views of the same information are often irreconcilable

- Global view does not reflect local information
  - **Simpson's paradox** in behavioral data
  - Global (population-level) trends may not reflect local (individual-level) tendencies
- Local views do not reflect the global reality
  - **Friendship paradoxes**
  - Network structure skews local perceptions of nodes

- What is Simpson’s paradox
- Why it occurs
- Some real-world examples
- How to test for it
- How to find it in data

Simpson's Paradox

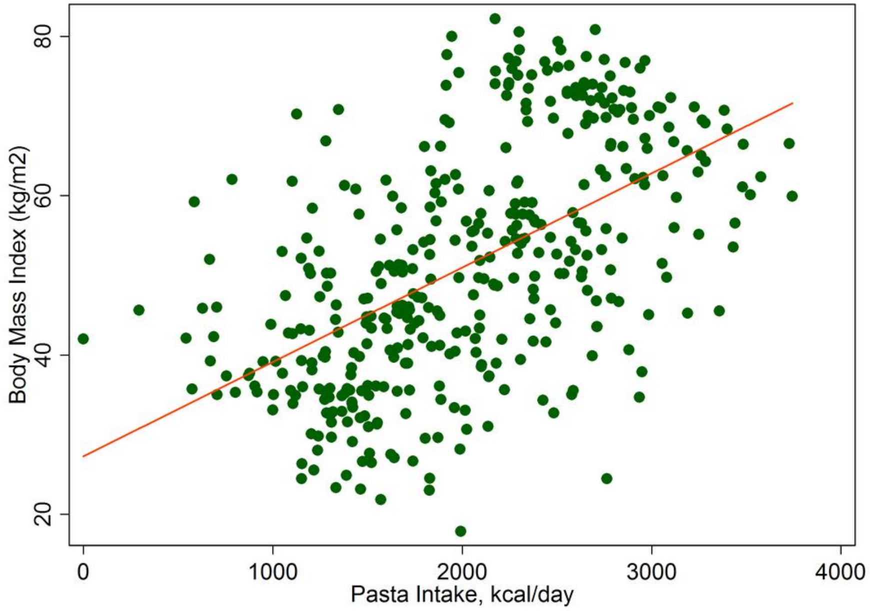
- A trend exists in aggregate data but disappears or reverses when data is disaggregated by subgroups

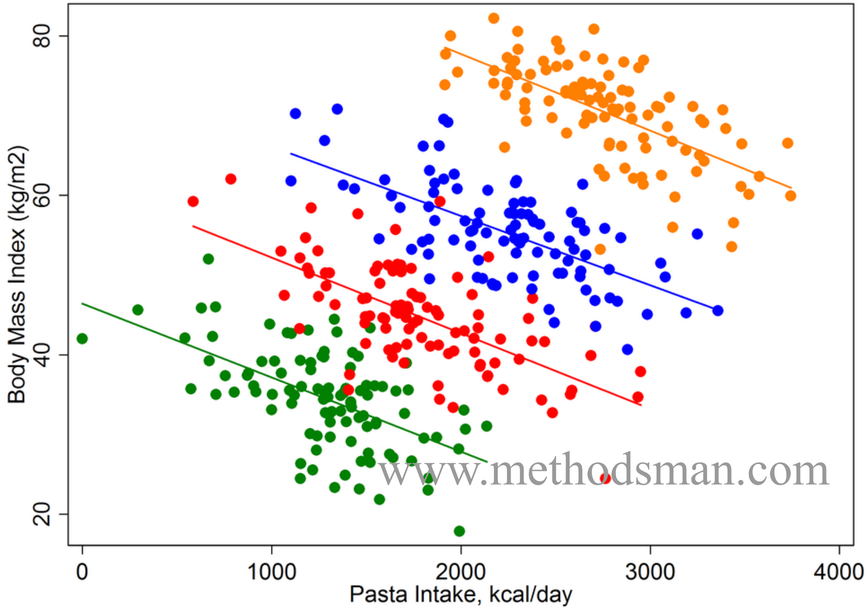


www.methodsman.com

SIMPSON'S PARADOX

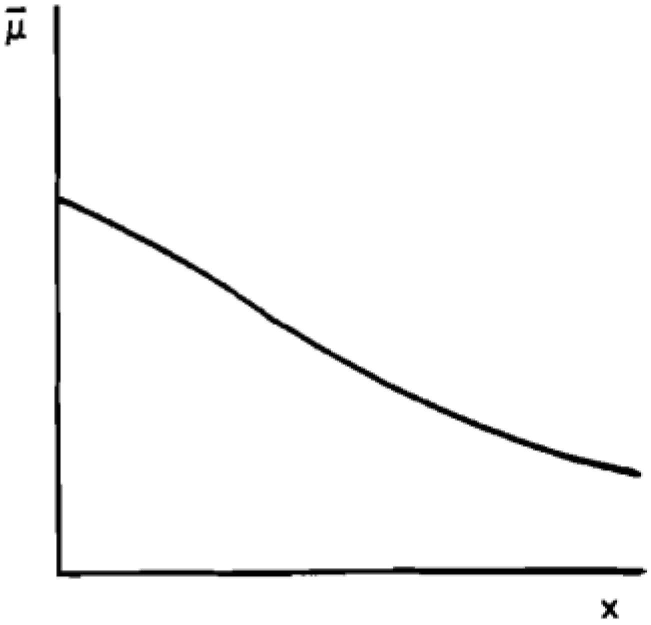
- A trend exists in aggregate data but disappears or reverses when data is disaggregated by subgroups





Survivor bias and heterogeneous population

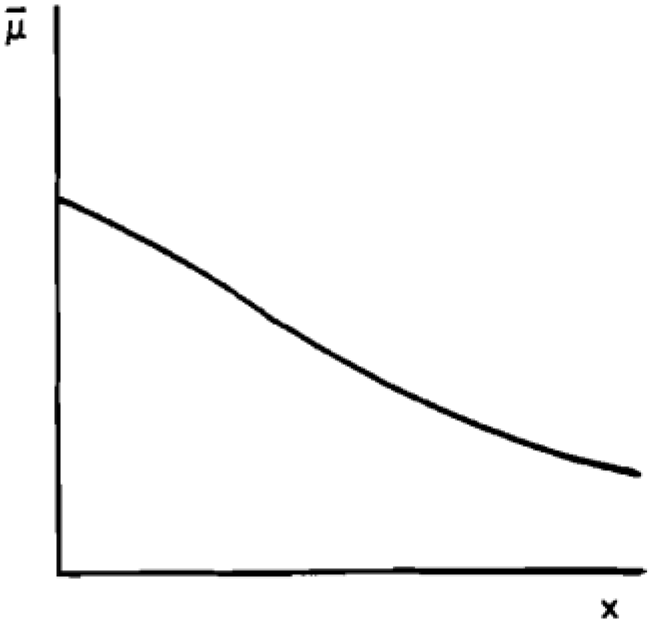
Recidivism rate of convicts released from prison declines with time since release



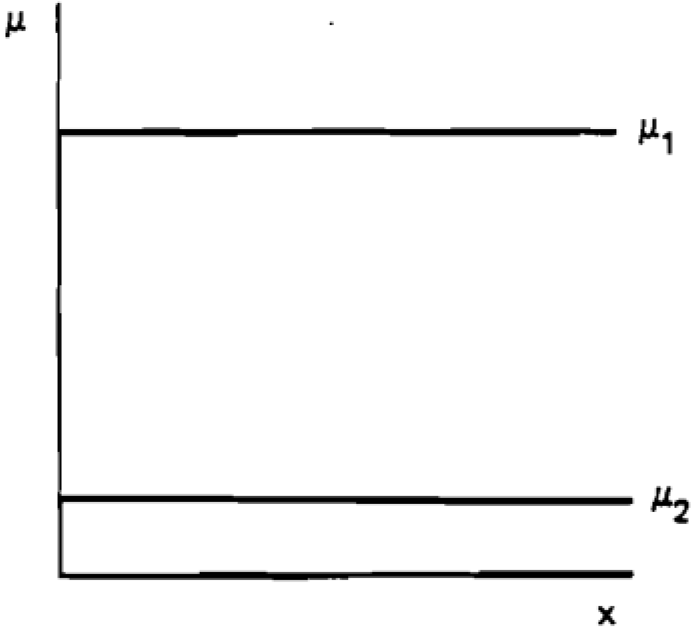
Vaupel, J. W. and Yashin, A. I. (1985). Heterogeneity's ruses: some surprising effects of selection on population dynamics. *The American Statistician*, 39(3):176-185.

Survivor bias and heterogeneous population

Recidivism rate of convicts released from prison declines with time since release



In reality, two populations: incorrigibles and reformed. Over time, fewer incorrigibles in the population



Vaupel, J. W. and Yashin, A. I. (1985). Heterogeneity's ruses: some surprising effects of selection on population dynamics. *The American Statistician*, 39(3):176-185.

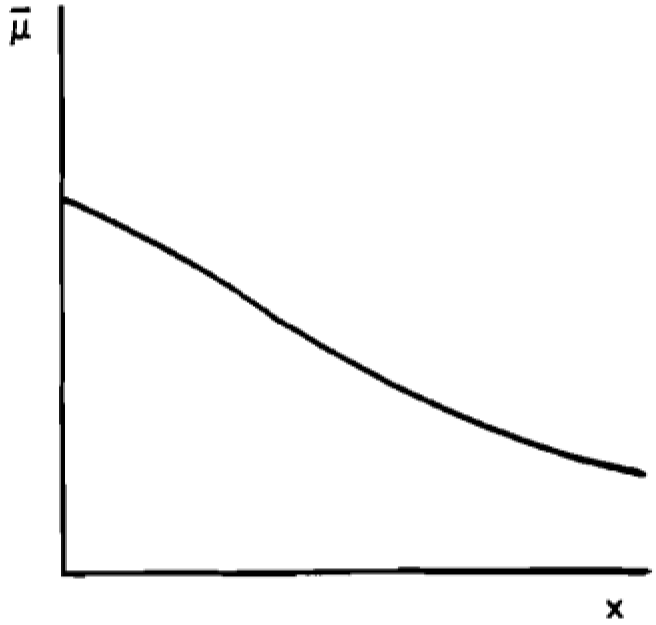
Why does Simpson’s paradox occur?

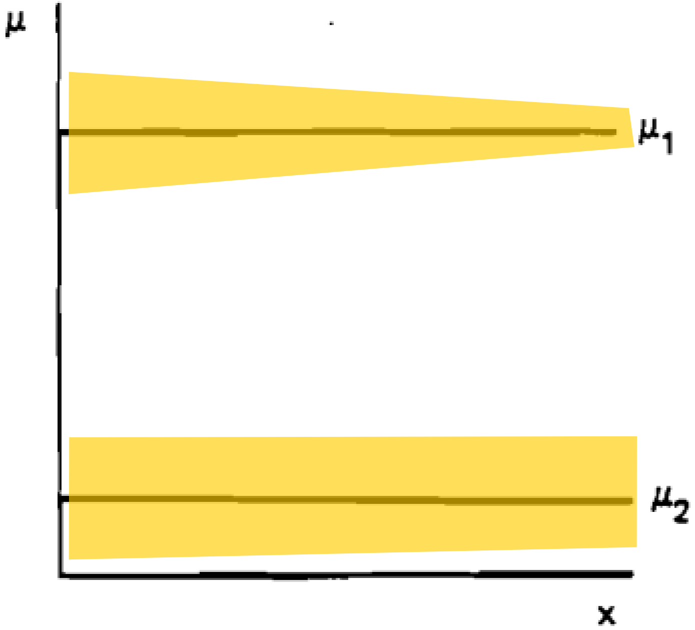
- Subgroups differ in the background factor
- The background factor and the independent variable are correlated

Survivor bias and heterogeneous population

Average rate appears to decrease…

… over time, there are fewer people from subgroup1 (incorrigibles) in the population



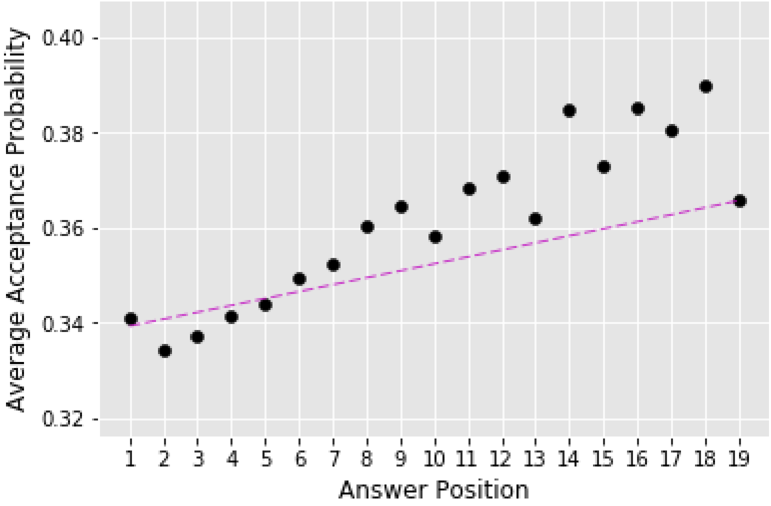


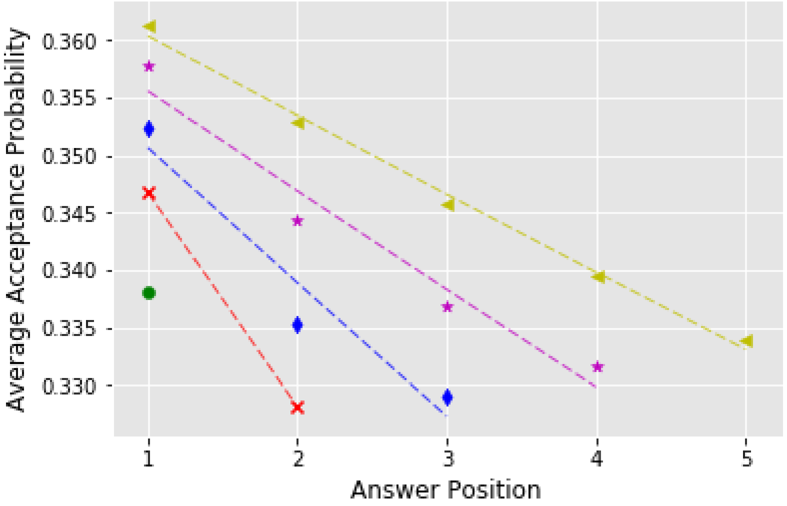
Vaupel, J. W. and Yashin, A. I. (1985). Heterogeneity's ruses: some surprising effects of selection on population dynamics. *The American Statistician*, 39(3):176-185.

Stack Exchange: deterioration in answer quality

Better answers?: Users appear to write better answers (more likely to be accepted as best answer) later in a session

Worse answers: When the same data is disaggregated by length of the session, later answers are less likely to be accepted.



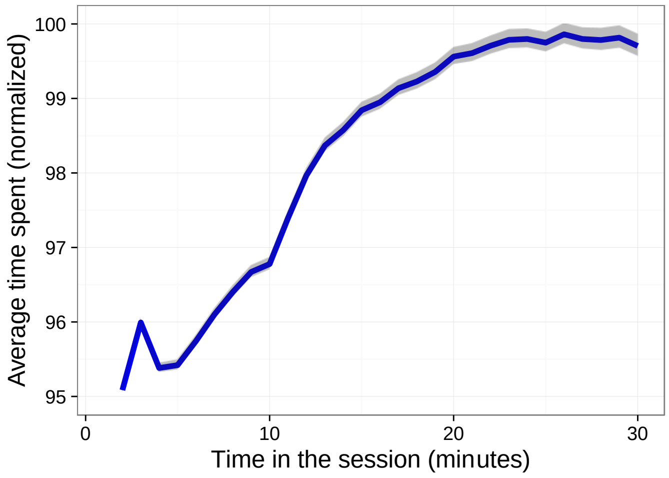


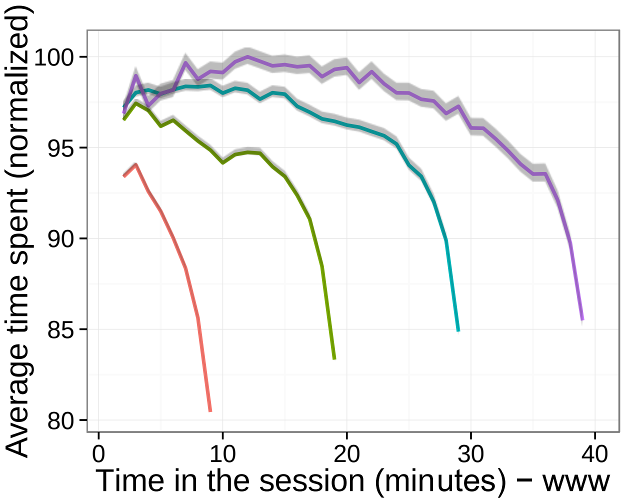
[Ferrara, Alipoufard, Burghardt, Gopal & Lerman (2017) "Dynamics of content quality in collaborative knowledge production", in *ICWSM*.]

Facebook: content consumption rates

Slowdown?: Facebook users appear to spend more time reading each story over the course of a session

Speedup: When the data is disaggregated by session length, users spend less time reading each story later in a session

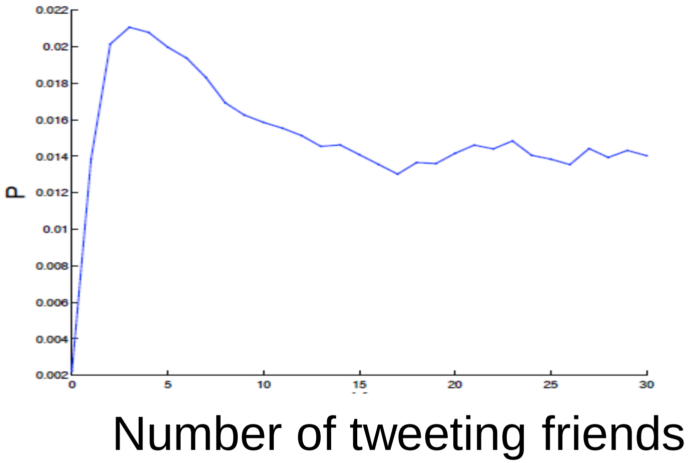




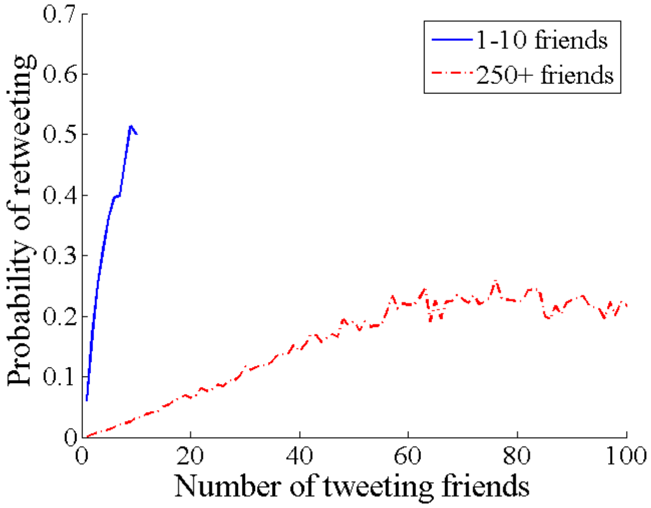
[Kooti, Subbian, Mason, Adamic & Lerman (2017) "Understanding short-term changes in online activity sessions", in *WWW*.]

Social contagion: do friends amplify or suppress response?

Complex contagion?: Additional exposures by friends appear to suppress response (probability to use a hashtag)[1]



Simple contagion?: When disaggregated by cognitive load (number of friends), additional exposures by friends amplify response (probability to retweet)[2]

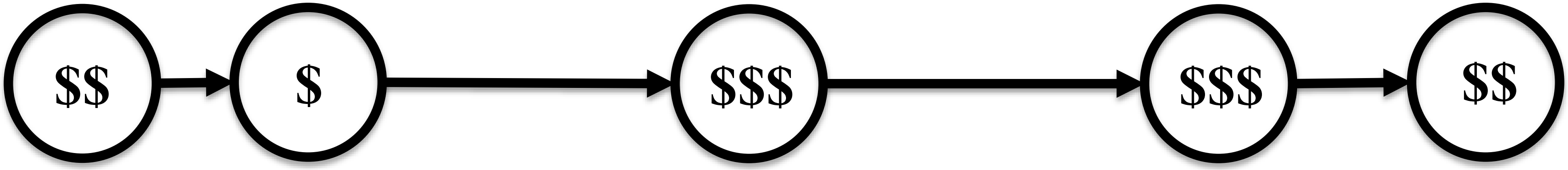


[1. Romero, Meeder & Kleinberg (2011) "Differences in the Mechanics of Information Diffusion Across Topics" in *WWW*.]

[2. Hodas & Lerman (2012) "How visibility and divided attention constrain social contagion", in *SocialCom*.]

How to test for Simpson’s paradox

The shuffle test

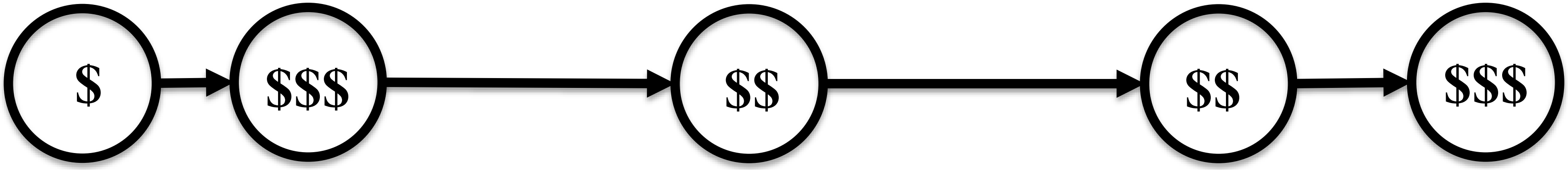


Randomize the data with respect to independent variable

- Trend should disappear in shuffled data
- E.g., online shopping: Is there a relationships between item price and how long a user waits to buy it?
  - Randomize the time items were purchased

[Lerman, K. (2018). Computational social scientist beware: Simpson's paradox in behavioral data. *Journal of Computational Social Sciences*, 1(1):49-58.]

The shuffle test



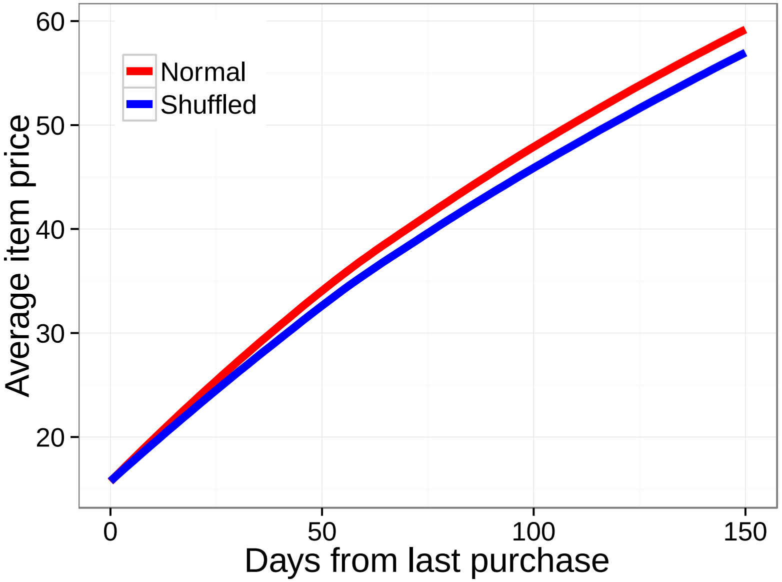
Randomize the data with respect to independent variable

- Trend should disappear in shuffled data
- E.g., online shopping: Is there a relationships between item price and how long a user waits to buy it?
  - Randomize the time items were purchased

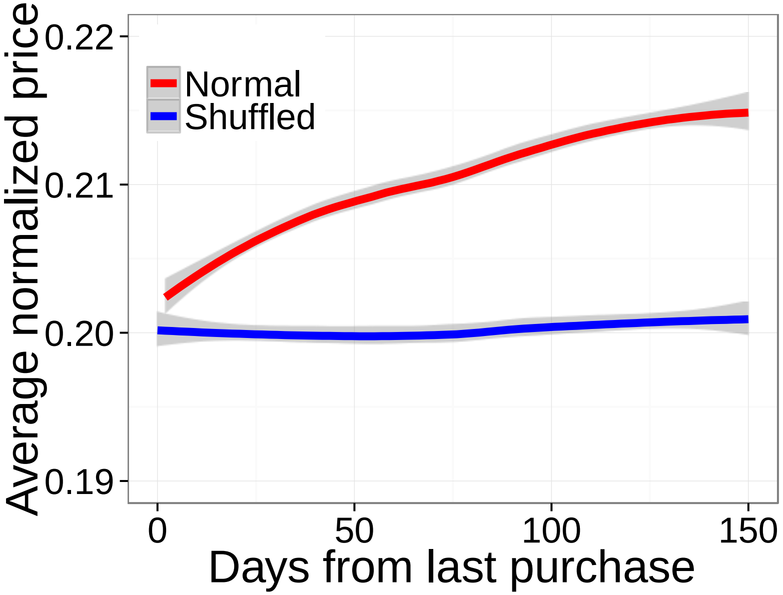
[Lerman, K. (2018). Computational social scientist beware: Simpson's paradox in behavioral data. *Journal of Computational Social Sciences*, 1(1):49-58.]

Testing the trend: online shopping

Online shopping: trend persists in the aggregated data after shuffling

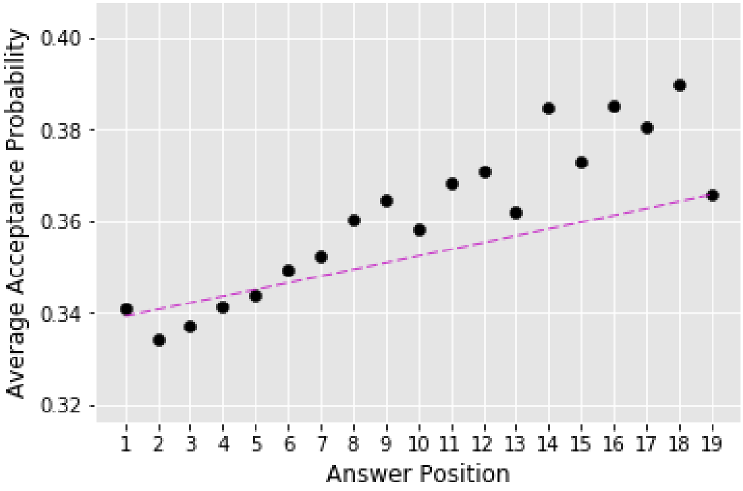


Online shopping: trend disappears (as expected) in the disaggregated data after shuffling

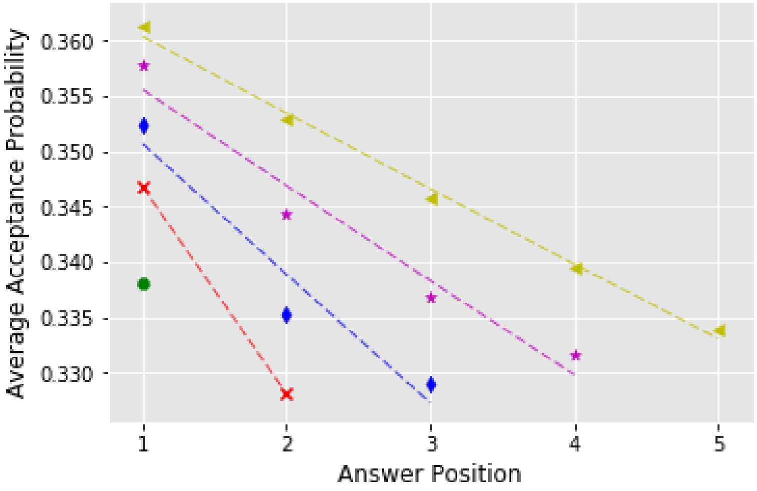


Users with 5 purchases

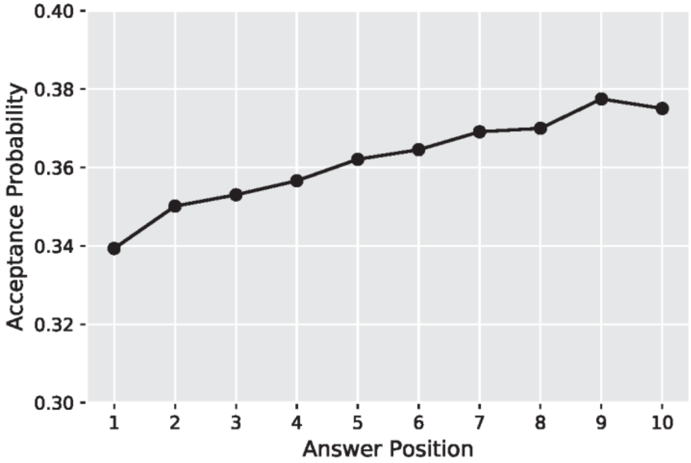
Stack Exchange: Original aggregate data



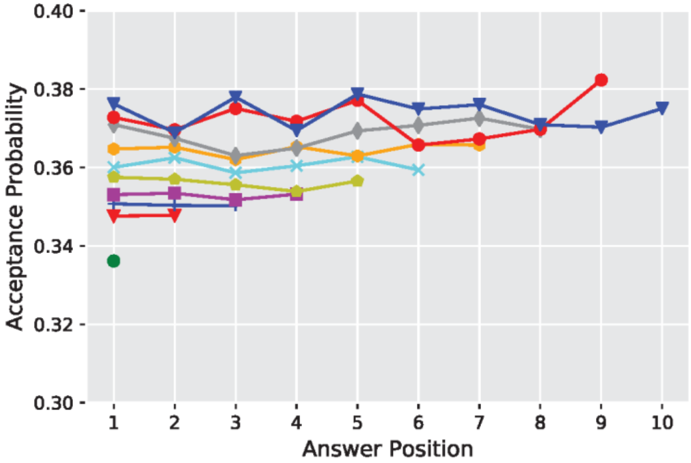
Original disaggregated data



Trend remains in the shuffled aggregate data

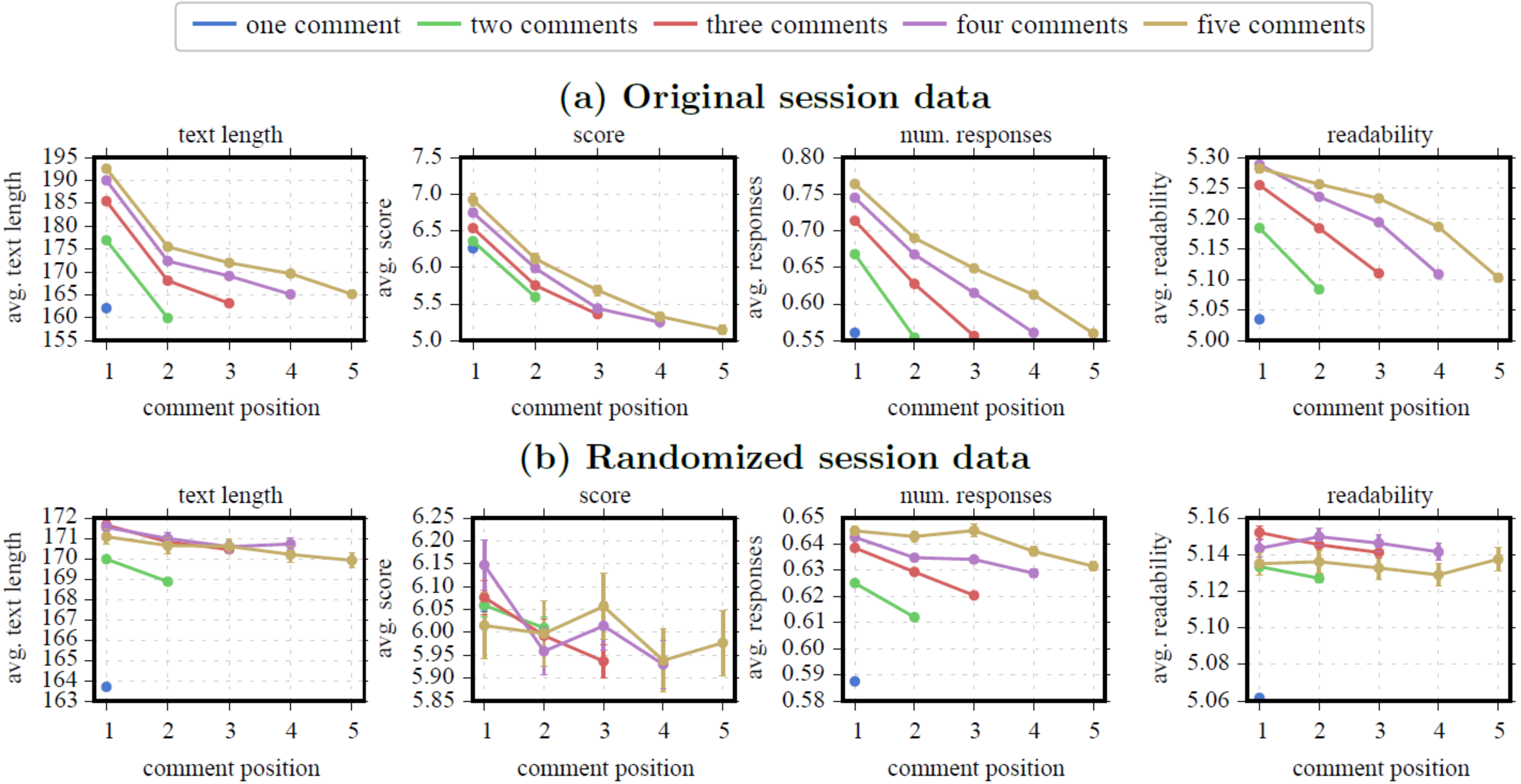


Trends disappear in the shuffled disaggregated data



[Ferrara, Alipoufard, Burghardt, Gopal & Lerman (2017) “Dynamics of content quality in collaborative knowledge production”, in *ICWSM*.]

Deterioration in comment quality on Reddit



(a) Original session data

(b) Randomized session data

➔ The more time people spend online, the worse they perform

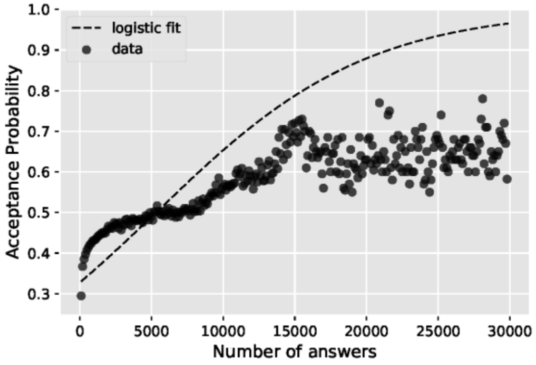
Automating discovery of Simpson’s paradoxes

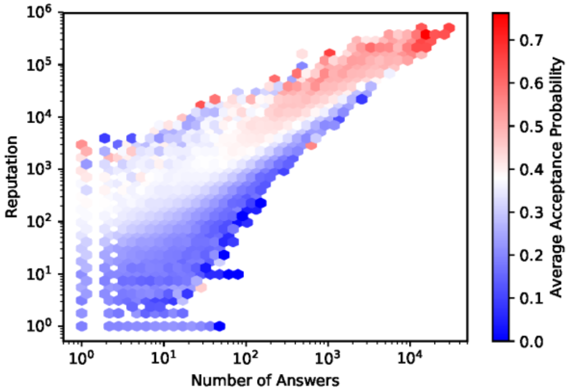
Method to discover Simpson’s paradoxes in data

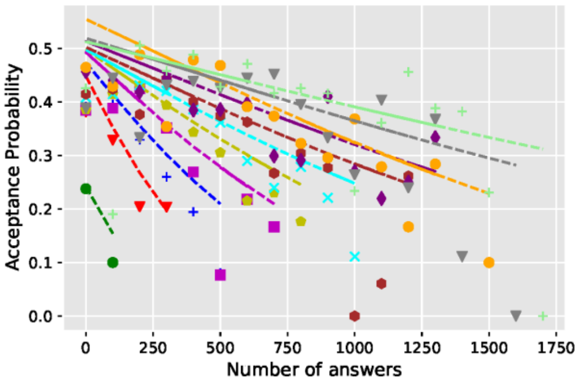
Step 1: Estimate trend with respect to an independent variable X_p

Step 2: Disaggregate data by conditioning on another variable X_c

Step 3: Compare trends in disaggregated subgroups to trends in aggregate data

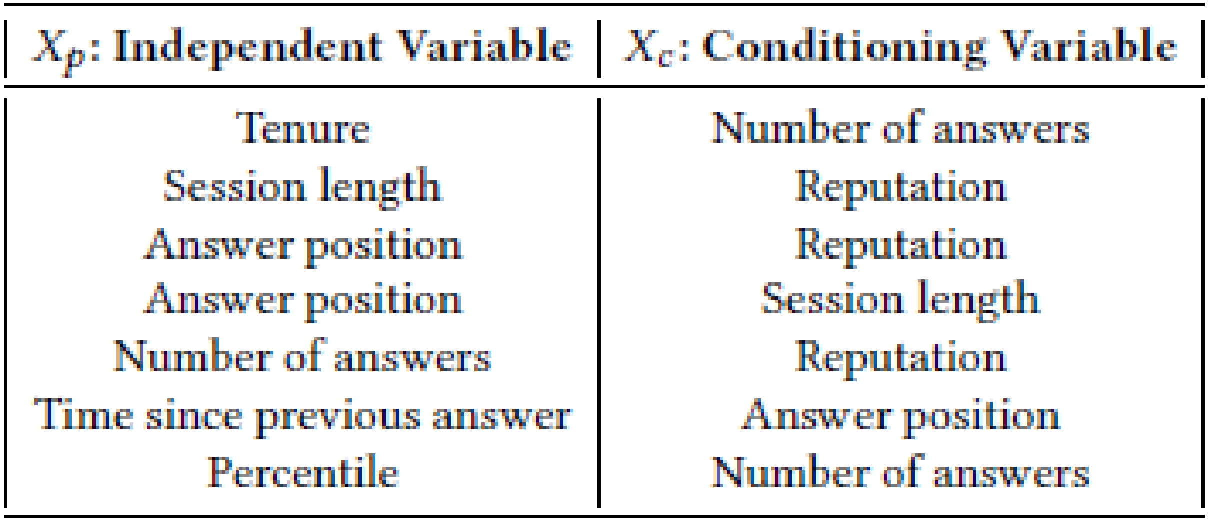






[Alipourfard, Fennell & Lerman (2017) “Don’t trust the trend: Discovering Simpson’s paradoxes in social data”, in WSDM.]

Paradoxes discovered in Stack Exchange data



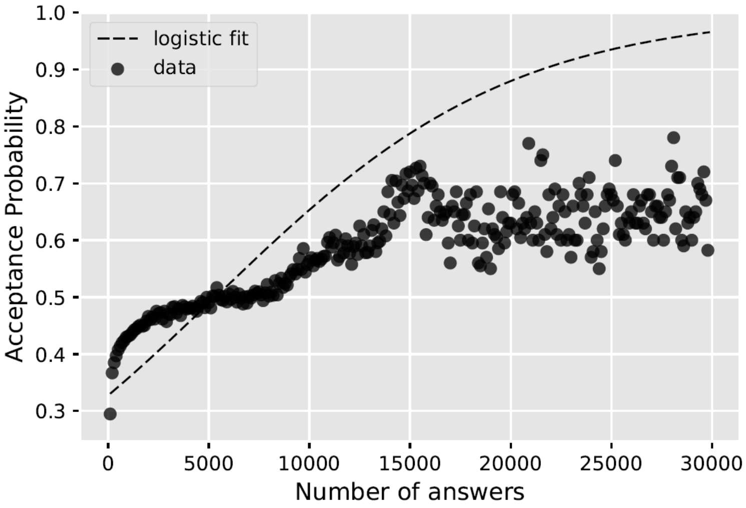
X_p: Independent Variable	X_c: Conditioning Variable
Tenure	Number of answers
Session length	Reputation
Answer position	Reputation
Answer position	Session length
Number of answers	Reputation
Time since previous answer	Answer position
Percentile	Number of answers

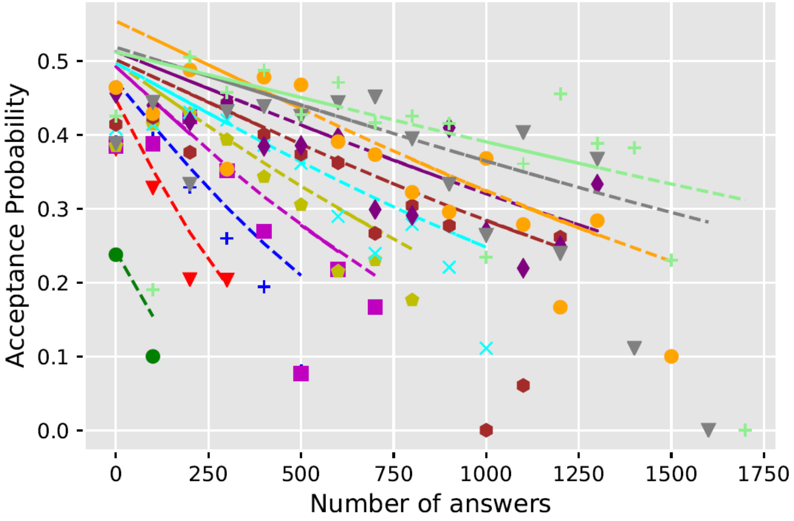
[Alipourfard, Fennell & Lerman (2017) “Don’t trust the trend: Discovering Simpson’s paradoxes in social data”, in WSDM.]

Stack Exchange: a new paradox we discovered

Does experience help?: Users who have already written more answers appear to write better answers (more likely to be accepted)

Worse answers: When the same data is disaggregated by reputation, having more experience does not help write better answers.

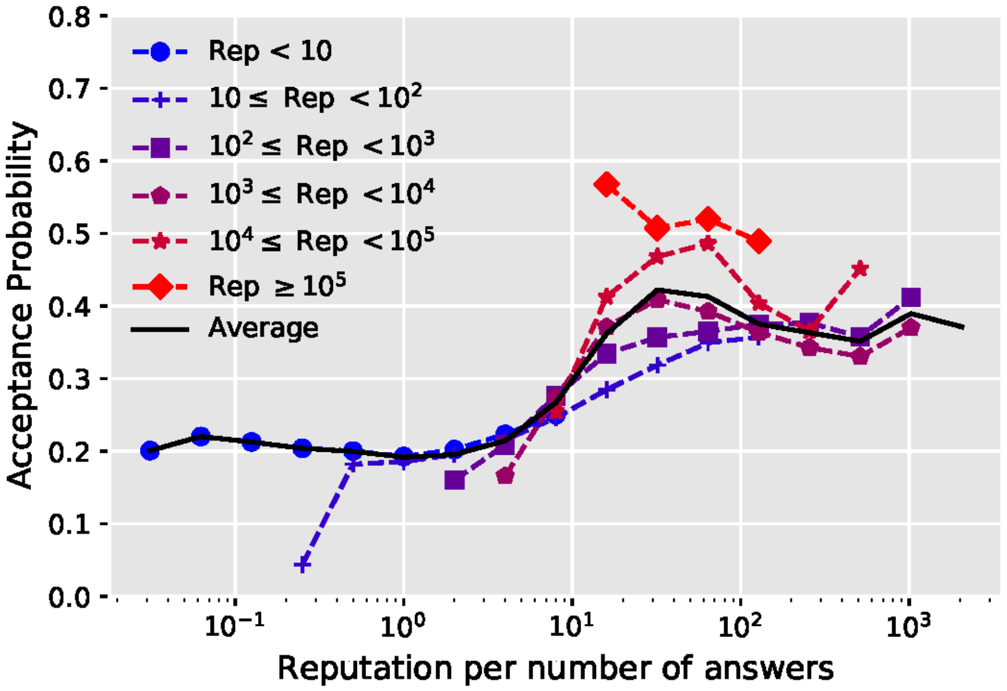




[Alipourfard, Fennell & Lerman (2017) “Don’t trust the trend: Discovering Simpson’s paradoxes in social data”, in WSDM.]

Data-driven discovery

Reputation Rate better explains behavior



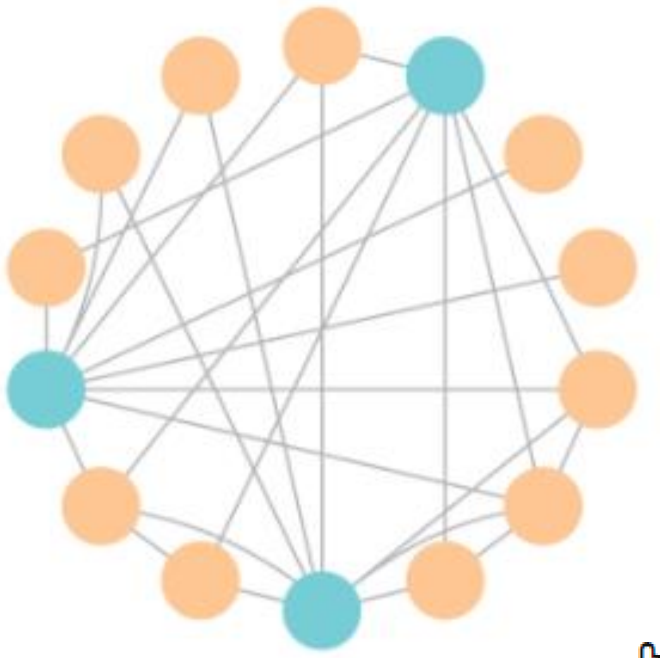
[Alipourfard, Fennell & Lerman (2017) "Don't trust the trend: Discovering Simpson's paradoxes in social data", in WSDM.]

FRIENDSHIP (AND OTHER) PARADOXES IN NETWORKS

Networks distort individuals' perceptions

The Washington Post

By Kevin Schaul



they are fashionable.

they are not.

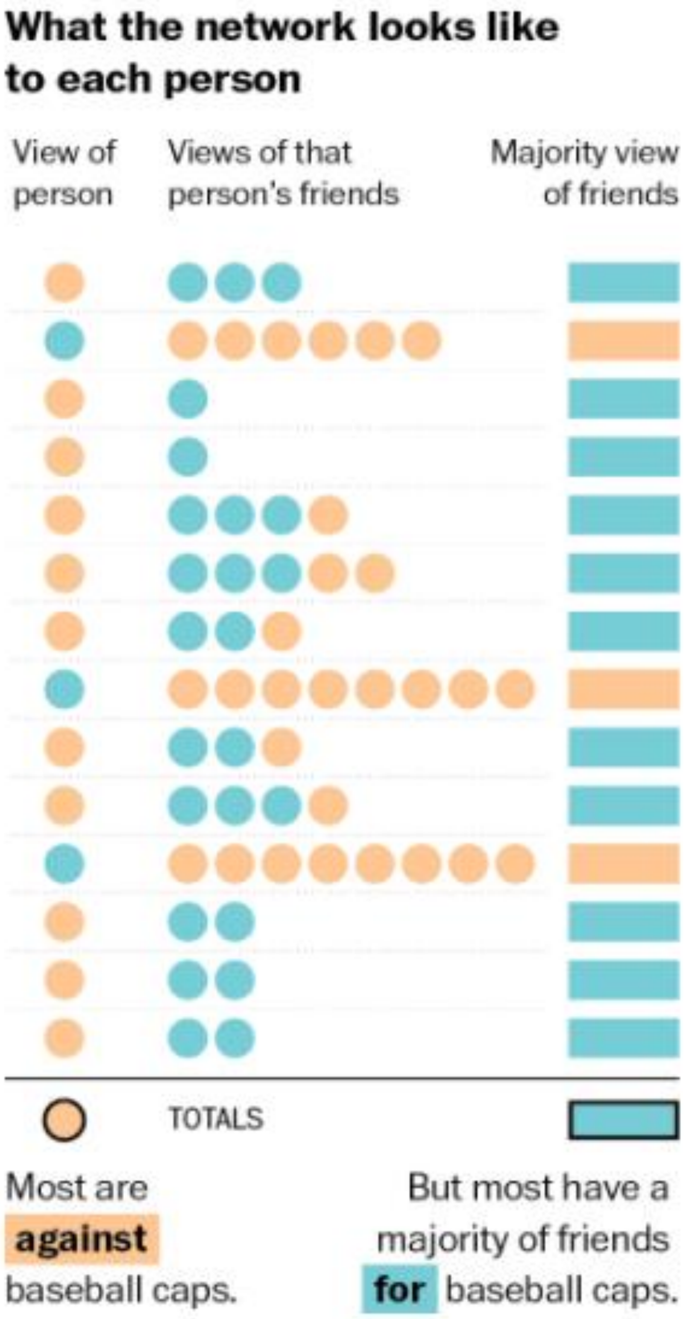
A town is voting to officially declare baseball caps fashionable. A polling firm asks people whether they thought baseball caps have popular support. People only know their own opinion and what their friends think.

☝ Did the polling firm find the measure was expected to pass or fail?

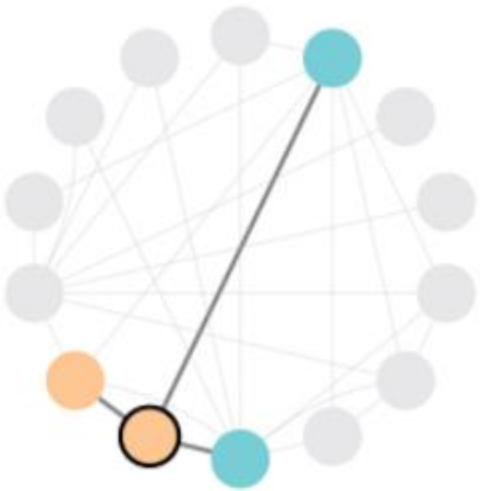
Pass **Fail**

Majority illusion

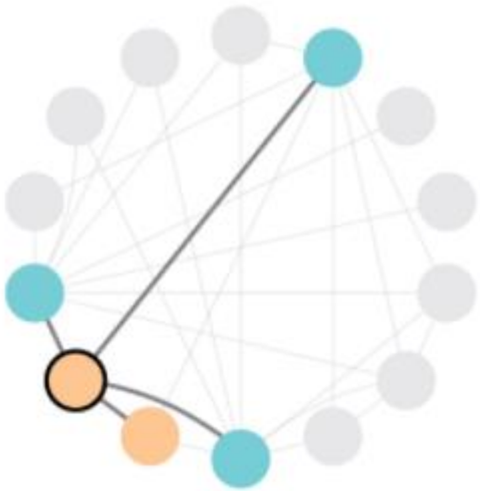
A minority opinion can appear to be very popular within many local social circles.



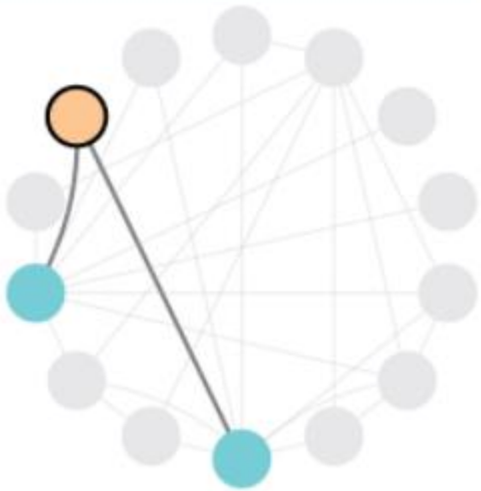
67% of this person's friends think baseball caps are trendy.



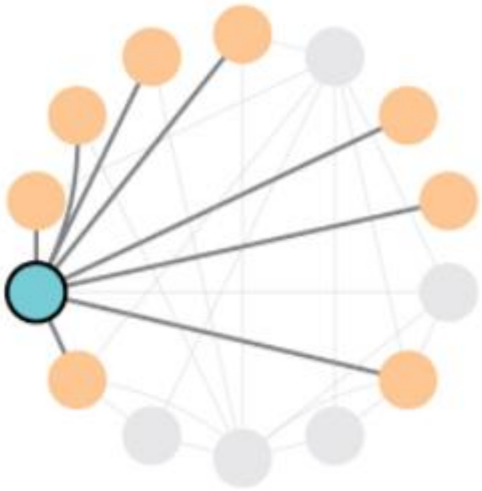
75% of this person's friends think baseball caps are trendy.



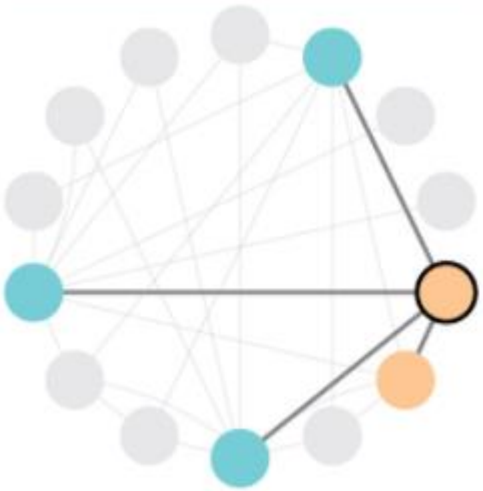
100% of this person's friends think baseball caps are trendy.



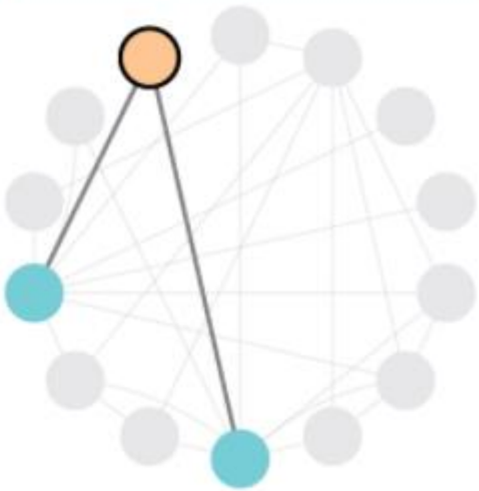
0% of this person's friends think baseball caps are trendy.



75% of this person's friends think baseball caps are trendy.

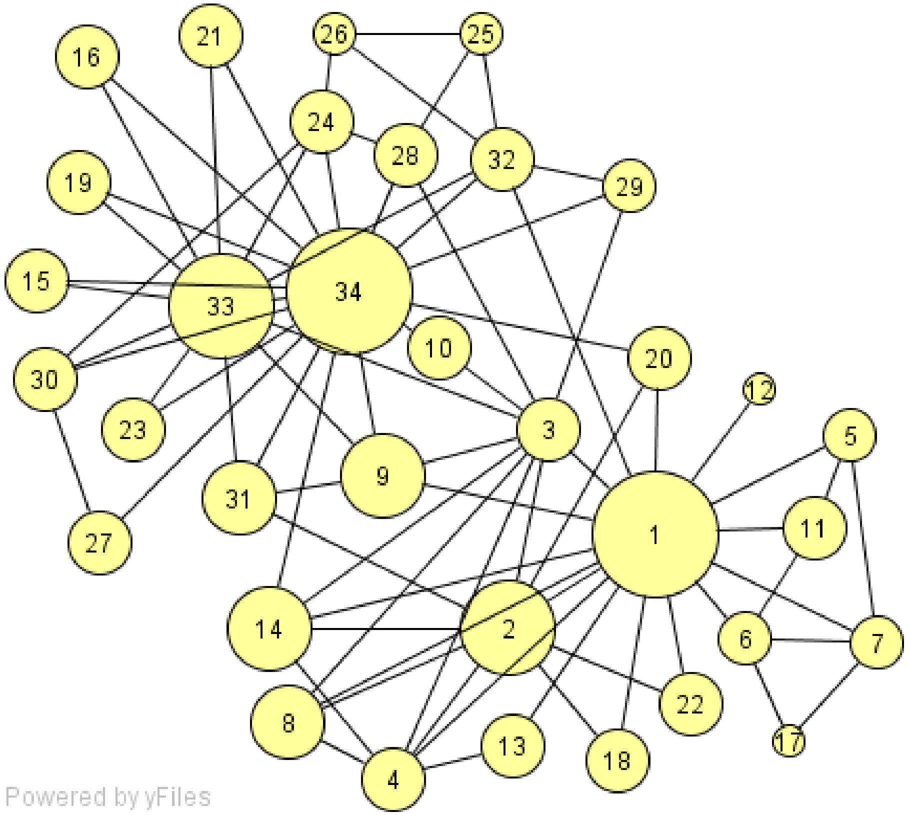


100% of this person's friends think baseball caps are trendy.



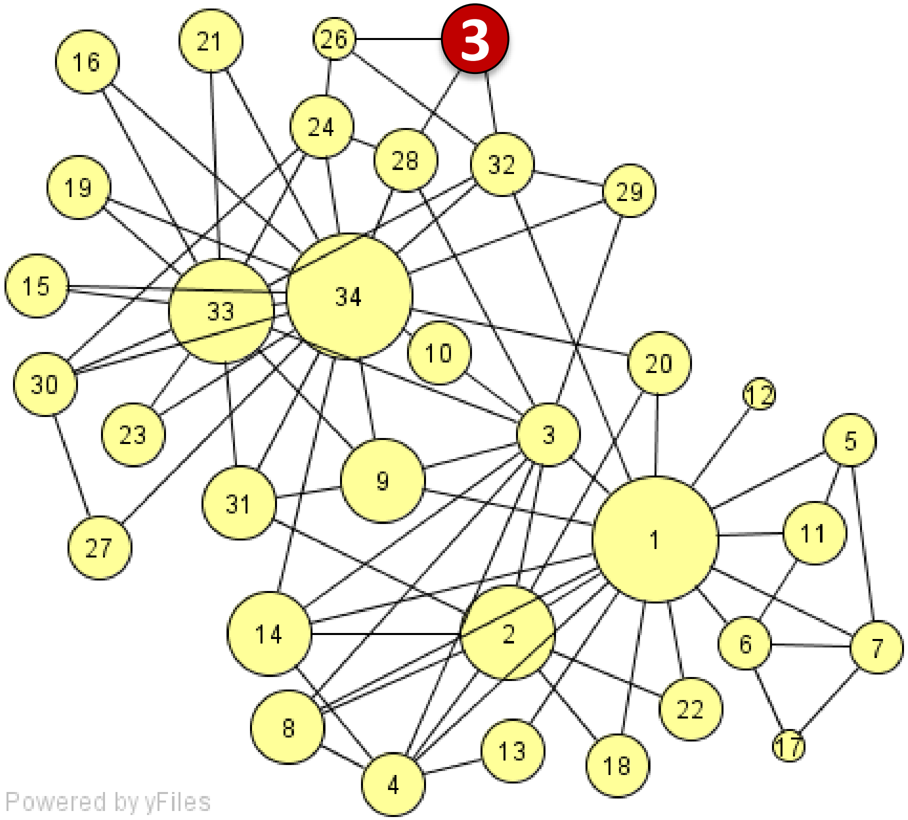
Friendship paradox

Friendship paradox: On **average**, your friends have more friends than you do [Feld, 1991].



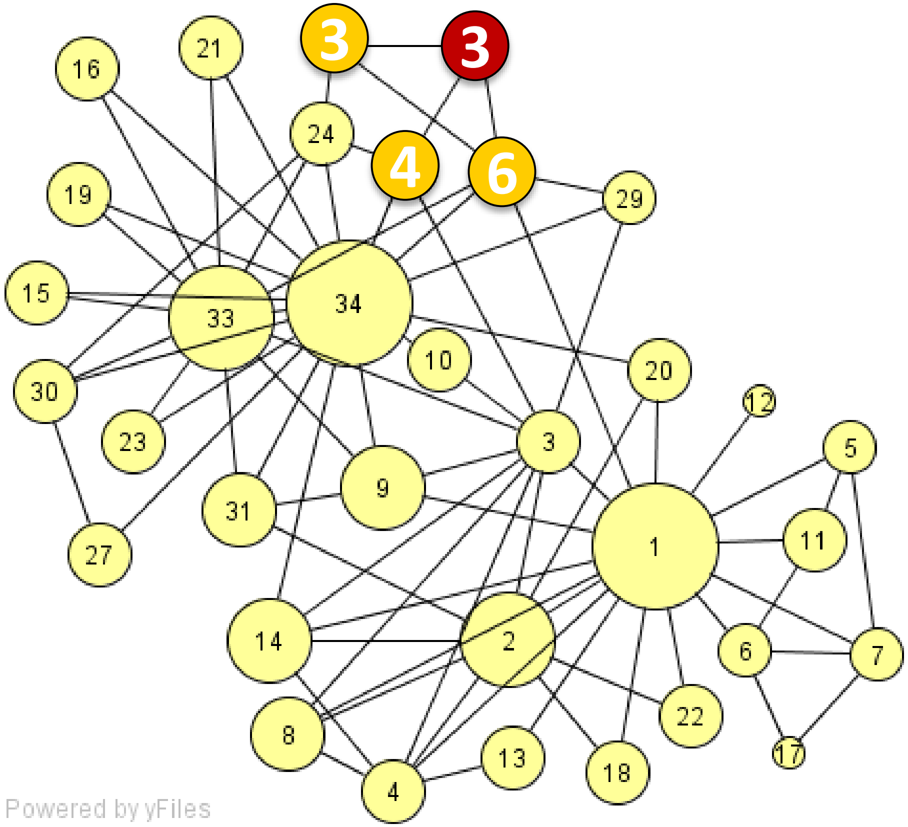
Friendship paradox

Friendship paradox: On **average**, your friends have more friends than you do [Feld, 1991].



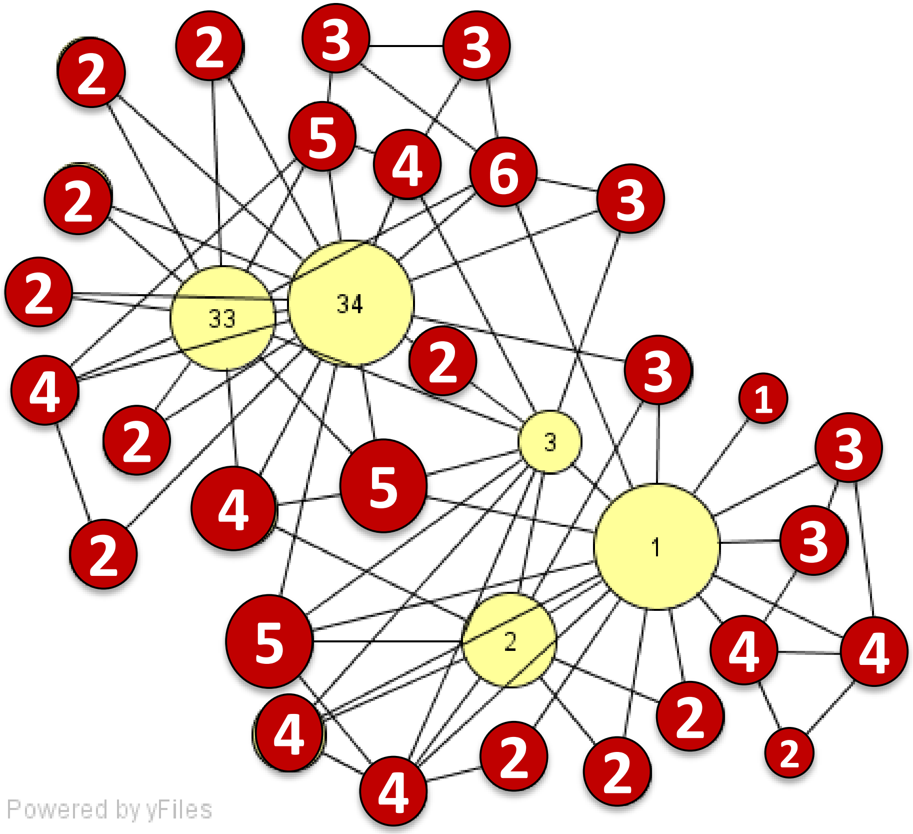
Friendship paradox

Friendship paradox: On **average**, your friends have more friends than you do [Feld, 1991].



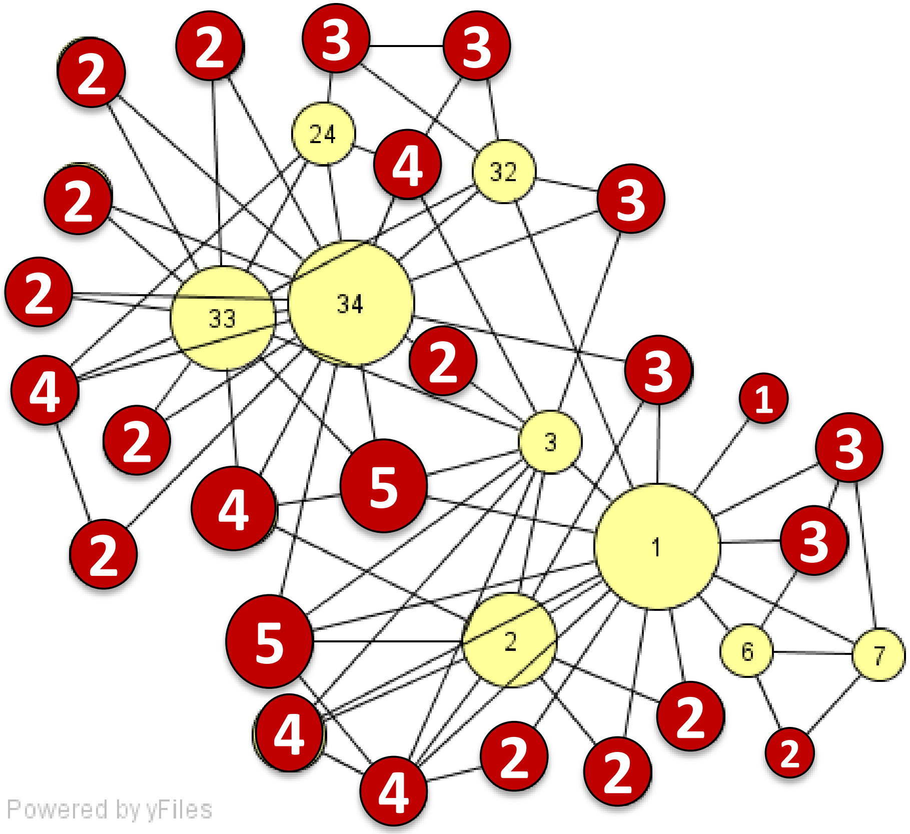
Friendship paradox

Friendship paradox: On **average**, your friends have more friends than you do [Feld, 1991].



Strong friendship paradox

Strong friendship paradox: Most of your friends have more friends than you do [Kooti, Hodas and Lerman, 2014].



How strong is strong friendship paradox?

A very large fraction of individual nodes observe that most of their neighbors have a larger degree

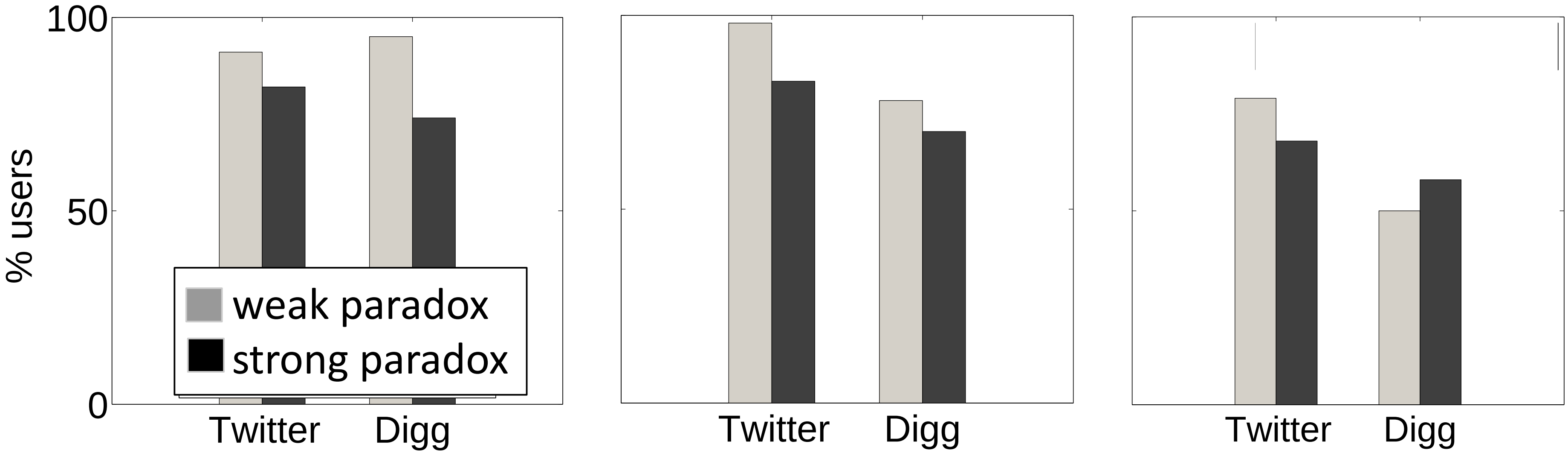
Network	Type	Nodes	Probability of paradox
LiveJournal	Social	3,997,962	84%
Twitter	Social	780,000	98%
Skitter	Internet	1,696,415	89%
Google	Hyperlink	875,713	77%
ProsperLoan	Social Finance	89,269	88%
ArXiv	Citation	34,546	79%
WordNet	Semantic	146,005	75%

Generalized friendship paradoxes

Activity paradox: Most of your friends post more messages than you do.

Diversity paradox: Most of your friends receive more diverse information than you do

Virality paradox: Most of your friends receive more viral information than you do.

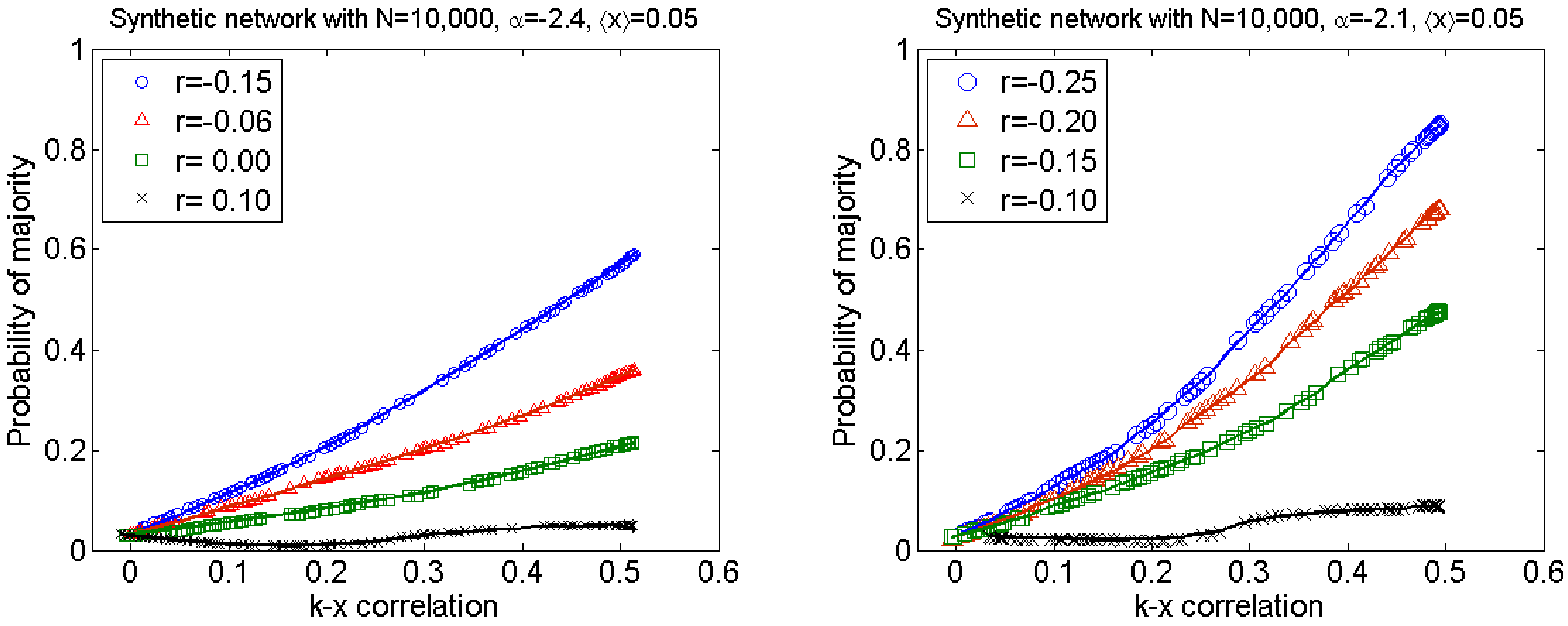


[Kooti, et al (2014) "Network Weirdness: Exploring the origins of network paradoxes" in *ICWSM*]

Strong friendship paradox creates majority illusion

When high degree nodes are more likely to have a trait, the remaining nodes will experience majority illusion

- Large degree-trait (k-x) correlation amplifies the illusion
- Stronger in disassorative networks (smaller r)

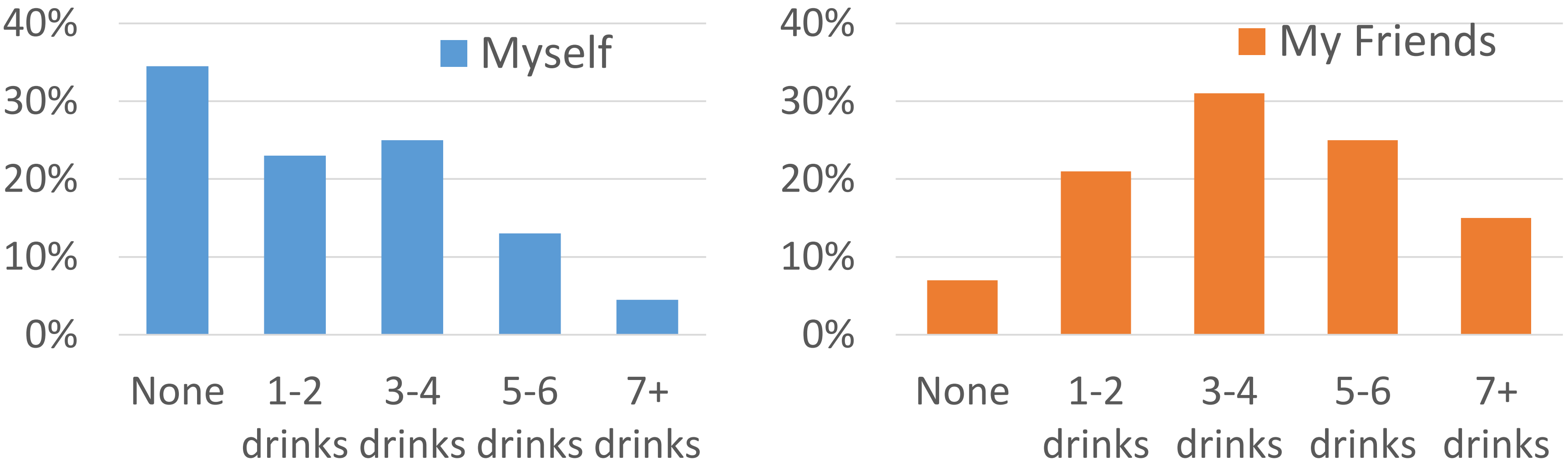


[Lerman, Wu & Yan (2016) The “Majority Illusion” in Social Networks, in *Plos One.*]

Friendship paradox and risky behavior

- Strong friendship paradox can systematically distort individual’s perceptions
- Example: College students overestimate peers’ alcohol use

How many alcoholic drinks are consumed at a party



Source: Most Students Do PartySafe@Cal

To summarize

- Network structure can systematically bias local perceptions
  - Heterogeneous degree distribution (1K structure)
    - Large inequality of connectivity
  - Disassortativity (2K structure)
    - Popular people linking to unpopular people
  - Neighbor assortativity (3K structure)
    - Degree correlation of neighbors
  - Degree-trait correlation
    - Popular people more likely to have the trait, e.g., be rich
- Open questions: What is the impact of network bias on
  - Collective dynamics in networks, e.g., contagious outbreaks
  - Network sampling and inference
  - Network control and intervention

To summarize

- Simpson’s paradox occurs when an association observed in the subgroups disappears or reverses when the subgroups are combined into one.
- Also occurs when measuring trends with respect to an independent variable
- Algorithm to automatically identify subgroups with different trends
  - A tool for data-driven discovery
  - And to formulate new hypotheses about data.

THANK YOU!

Sponsors

NSF: CIF-1217605

ARO: W911NF-15-1-0142, W911NF-16-1-0306

Questions?

lerman@isi.edu