PART III. Considerations for large-scale and network data

PART III.
Special considerations with large-scale and network data

- High-dimensional data
- Network Effects
- Data about People

High-dimensional data creates estimation problems

High-dimensional data is often sparse.

E.g., Text: Distributed over all possible words

E.g., Medical: Not all tests given to every patient

Common statistical problem, reduces overlap (no two units are identical).

1. Consider **dimensionality reduction** techniques such as PCA.
2. Use **regularized models** for estimating propensity or regression.
3. **Transform input space** to obtain more overlap.

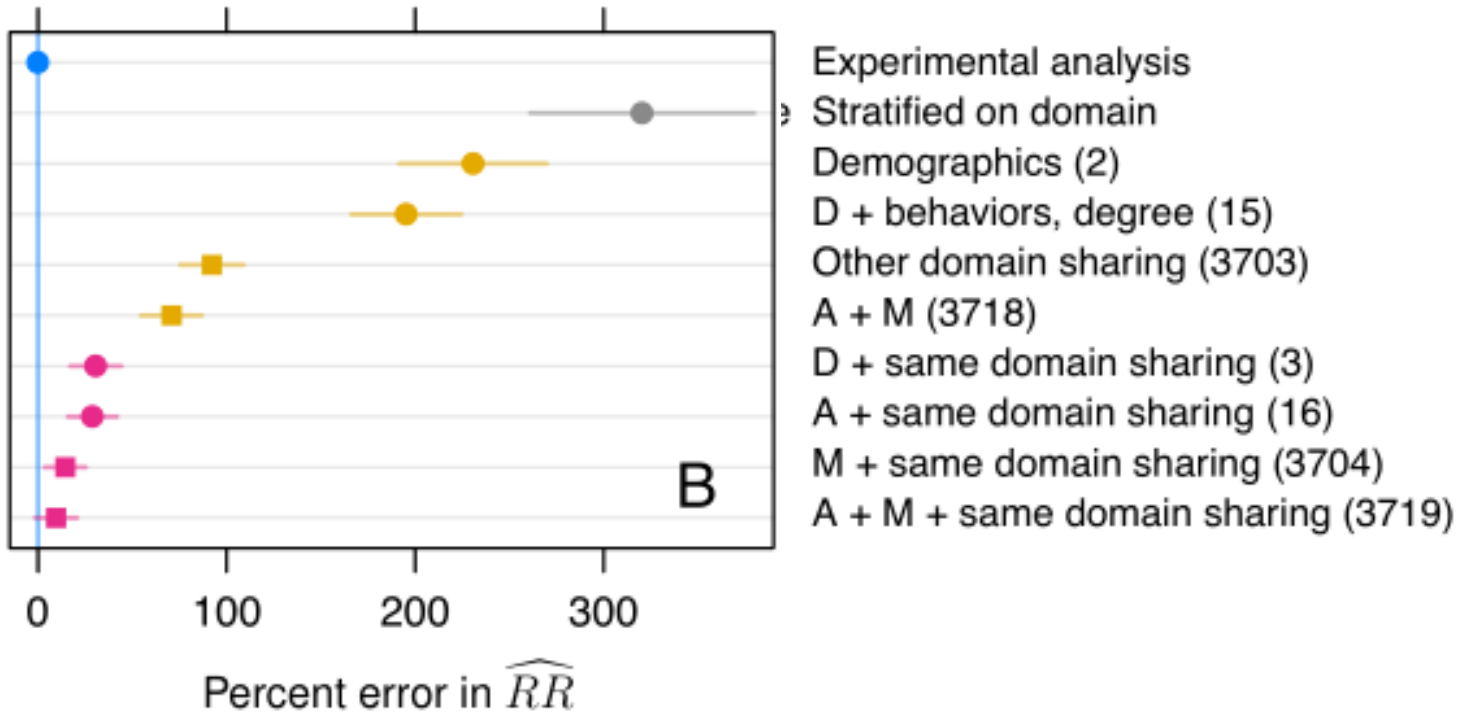
Example: Regularized propensity score

What is the effect of the Facebook news feed?

Counterfactual: Would a user have shared a URL had they not exposed to it using the Feed?

Over 3700 co-variates for matching.

Use logistic regression with L2 regularization.



Eckles and Bakshy. Bias and high-dimensional adjustment in observational studies of peer effects.

Paradox of measuring text with changing post frequency

Q: Should we use word counts or word probabilities when comparing text messages from two different groups?

If group A posts more frequently than group B then,

… any word w is used more frequently by A than by B. *Counts biased.*

… at the same time, $P_A(w) < P_B(w)$. *Likelihoods biased too.*

- Language models are hard to compare when vocabularies differ
  - Challenge: how to model and compare "out-of-vocabulary" likelihoods
- Mitigations:
  - Some heuristics for smoothing language models, but require tuning of OOV mass
  - Use language model over a fixed vocabulary (ignoring OOV)

PART III.
Special considerations with large-scale and network data

High-dimensional data

Network Effects

Data about People

Interference due to network effects

Network effects complicate causal inference.

If a person is exposed to some information, she might share with her friends.

Due to the exposure, her friends' outcome may also change.

Breaks the SUTVA assumption: an individual's outcome should not depend on another's treatment status.

1. Consider partitioned sub-networks as a unit of analysis.
2. Design alternative randomization assignment or estimator.

Example: Identifying peer effects with observational studies is impossible(!)

Consider the problem of separating peer influence from homophily.

Observed data: Activity of person i at time t is the same as activity of their friend j at time *t-1.*

Problem: Unobserved traits led to their friendship and also to the above common activity.

Without knowing all relevant latent traits, causal identification is impossible.

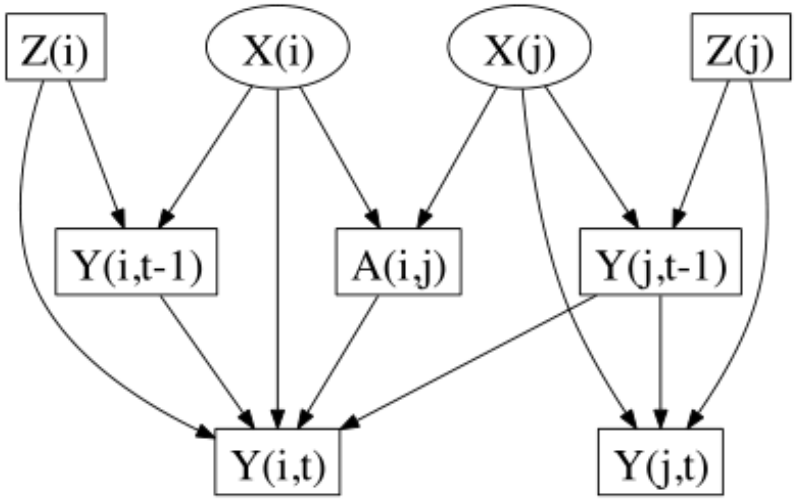


Figure 1: Causal graph allowing for latent variables (X) to influence both manifest network ties A_{ij} and manifest behaviors (Y).

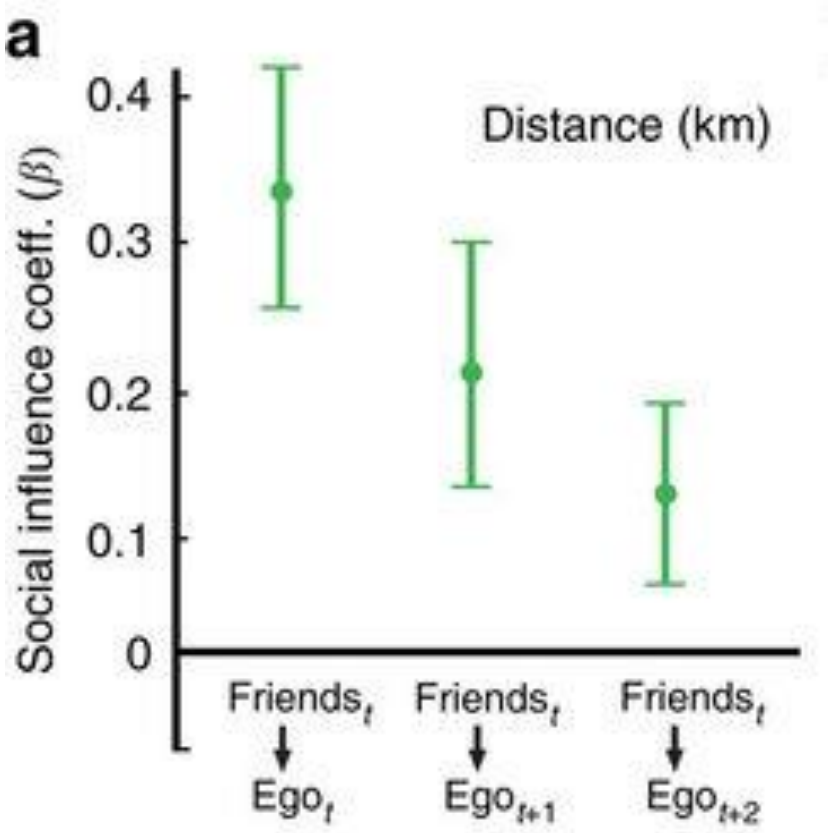
Shalizi and Thomas. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies.

Example: Use aggregated sub-networks

Question: Do peers influence us to exercise?

Instead of individuals, consider all users in a city as a unit.

Use rainfall to construct a natural experiment on running and do checks to validate IV assumptions to the extent possible.



Aral and Nicolaides. Exercise contagion in a global social network. Nature communications 2017.

PART III.
Special considerations with large-scale and network data

- High-dimensional data
- Network Effects
- Data about People

Everything depends on context

- Estimated effect is often context-dependent
  - May not generalize to other users
  - May not generalize to other platforms
  - May not generalize to other cultures

The WEIRD problem of social science studies.

1. Corroborate findings with multiple platforms or user samples.
2. Be explicit about plausible (non)-generalizability of causal effect.

Common confounders that lead to selection bias

Structured

Demographics (e.g., gender, age, income)

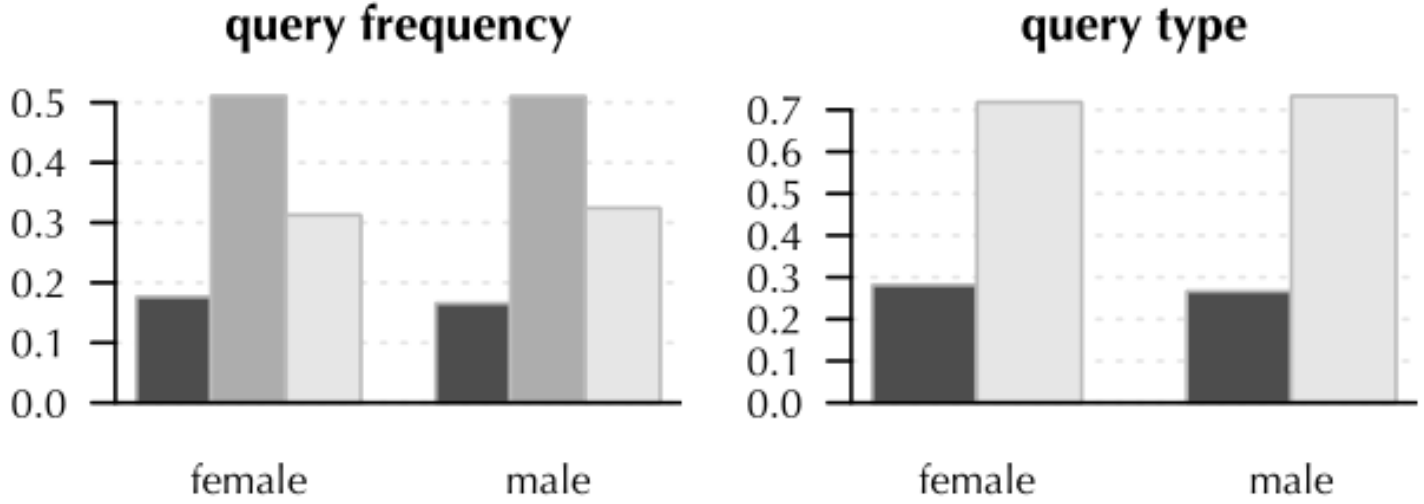
Patterns of usage(e.g., number of logins, type of activity)

Unstructured

Activity (e.g., post content, images)

Preferences (e.g., items interacted with)

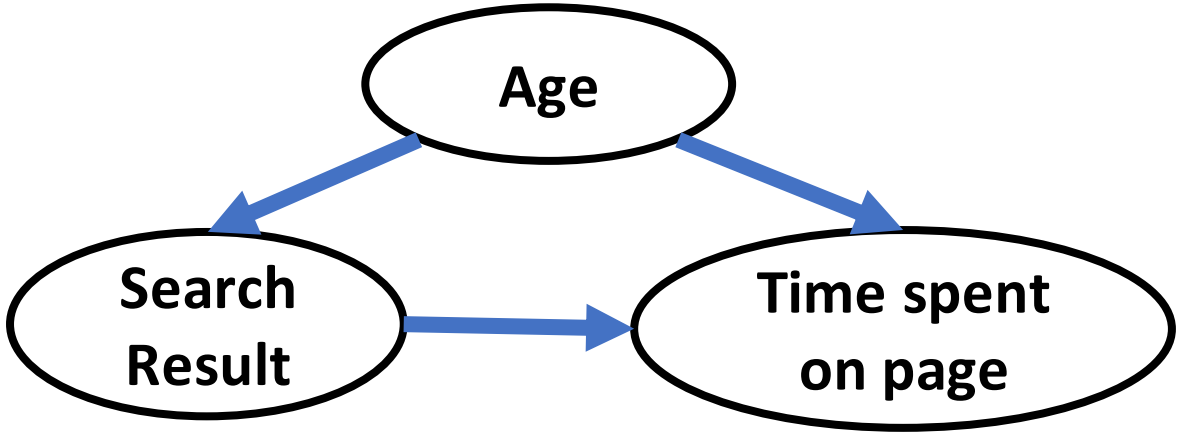
Demographic Bias



Online activity varies by demographics such as Age and Gender.

Search engines, recommendation feeds are measured on metrics such as "*Time spent on referred page*".

Without controlling for Age, metric is not trustworthy.

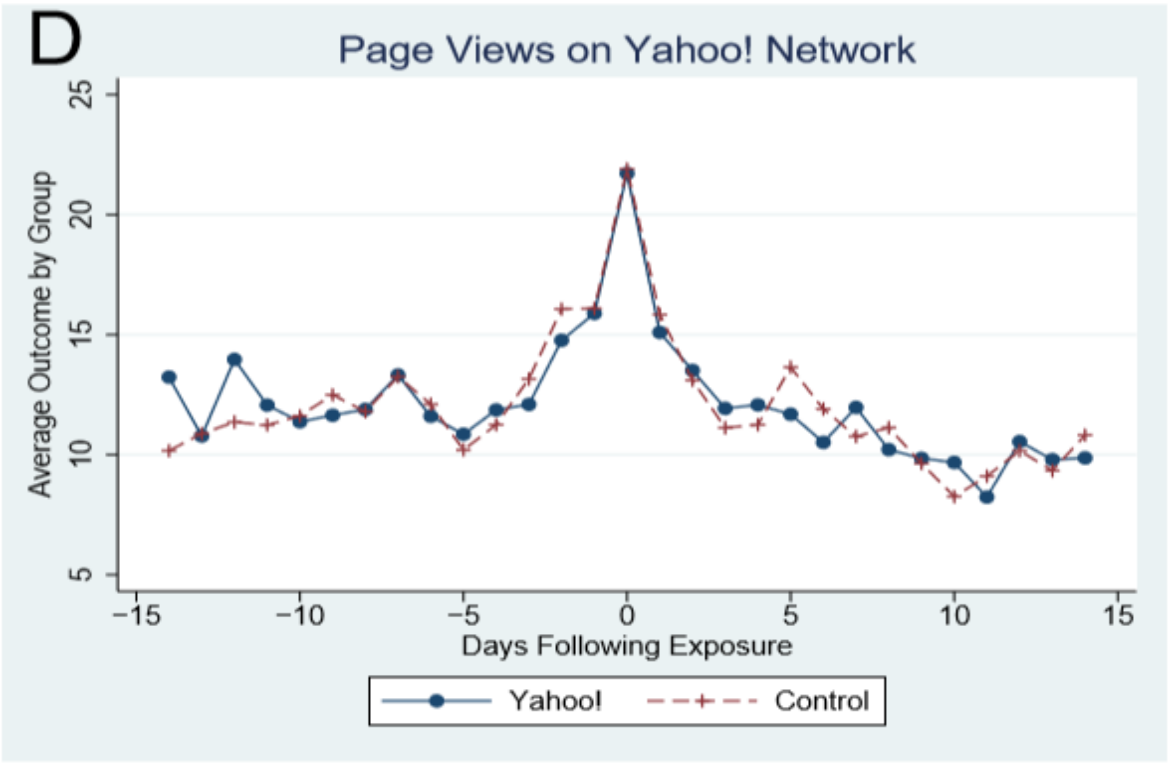


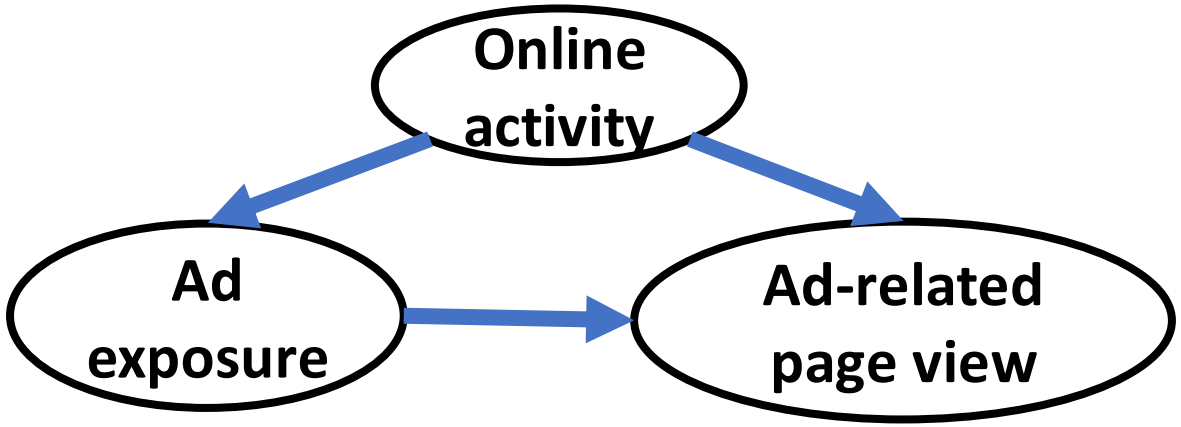
Mehrotra et al. Auditing Search Engines for Differential Satisfaction Across Demographics, WWW 2017.

Usage Bias

More activity can simply mean that people are online at the time, not due to any specific treatment.

People browse more ad-related products when they are shown an ad. But they also browse more of everything!





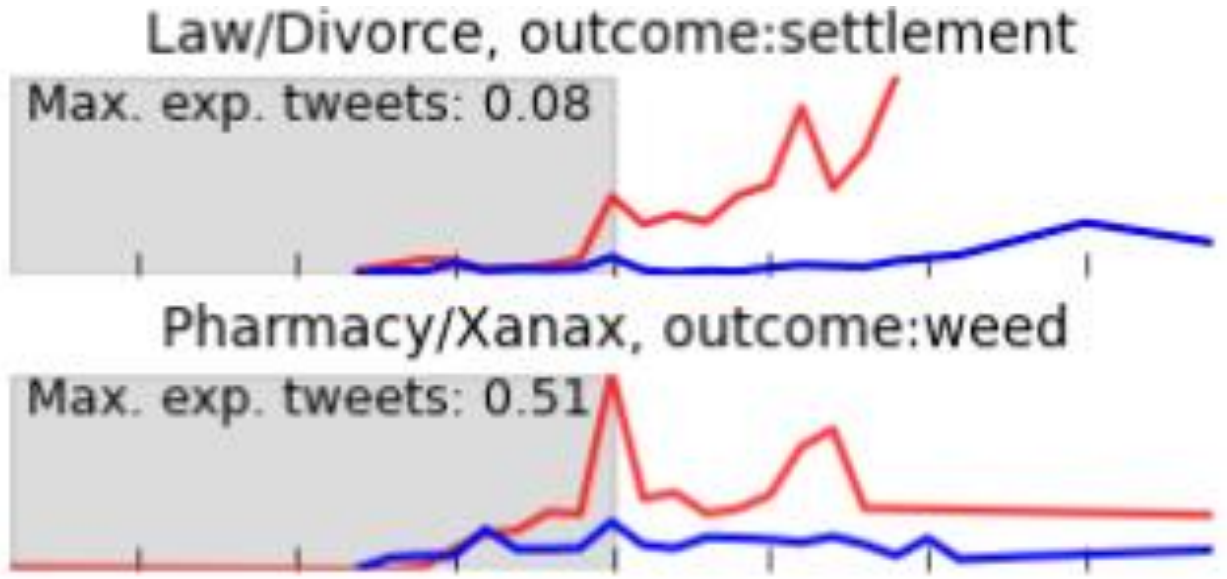
Lewis, Rao and Reily. Here, There, and Everywhere: Correlated Online Behaviors Can Lead to Overestimates of the Effects of Advertising. WWW 2011.

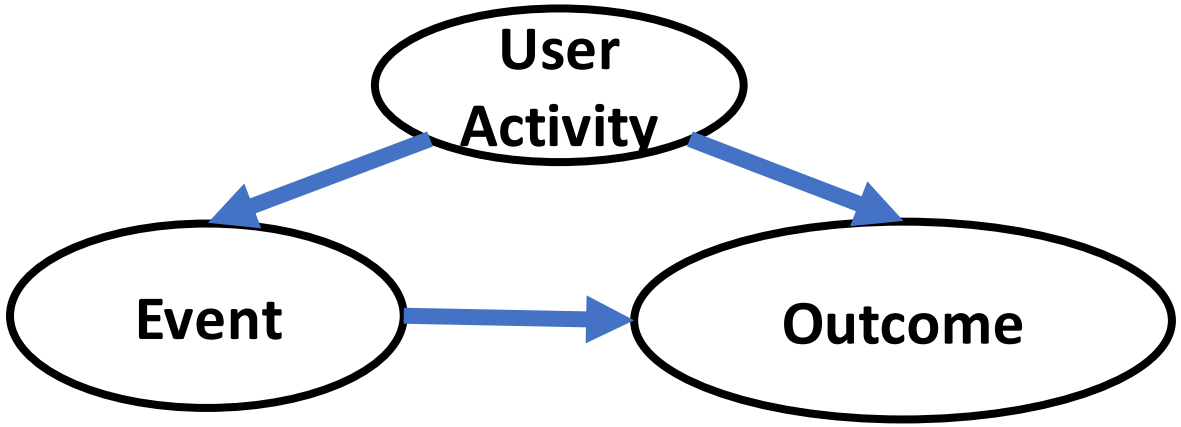
Activity Bias

Treated and untreated people may differ in many aspects.

People with demonstrably different activity content should not be compared.

Match people with similar activity content that is relevant to chances of being treated.





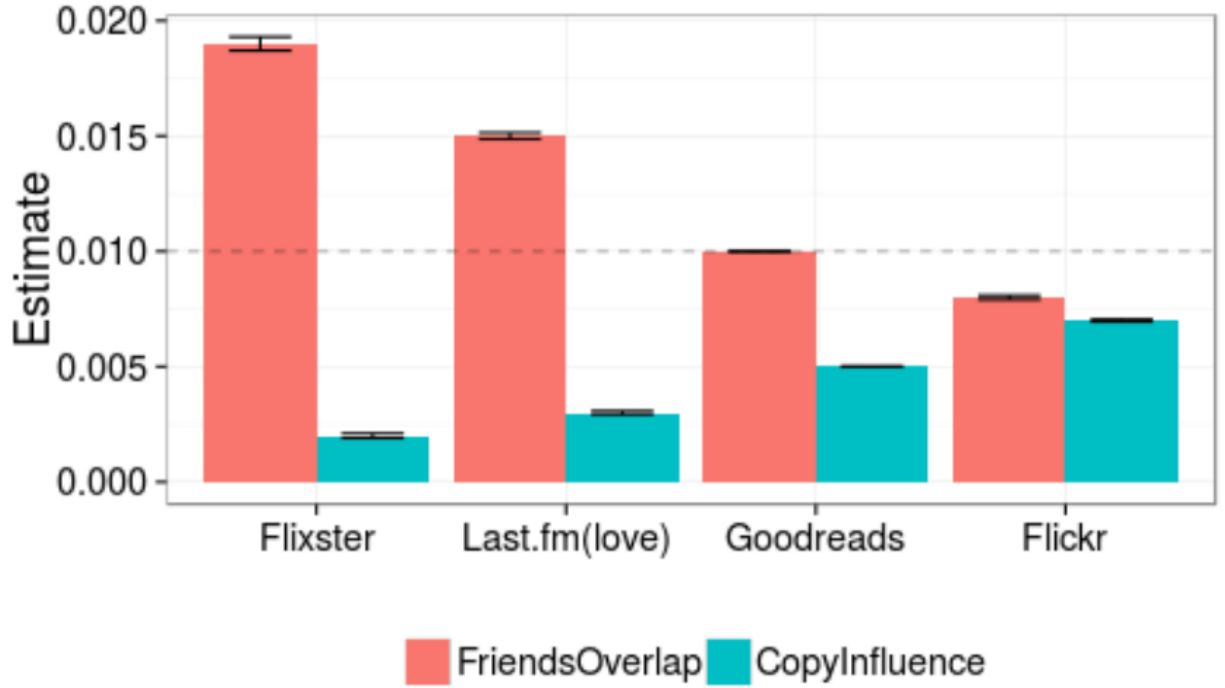
Olteanu, Varol and Kiciman. Distilling the Outcomes of Personal Experiences: A Propensity-scored Analysis of Social Media. CSCW 2017.

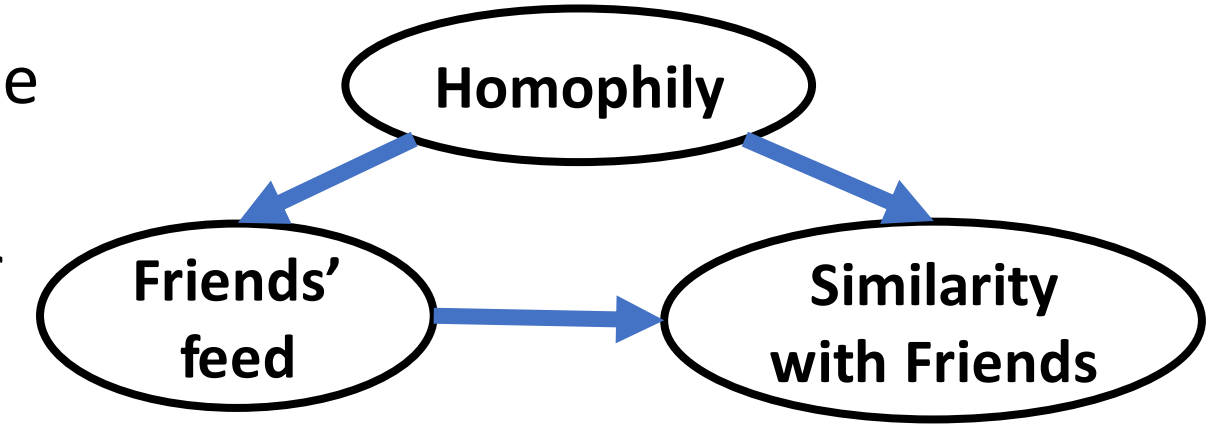
Preference Bias

Any similarity in activity may be due to inherent preferences, not any specific treatment.

Social influence from friends' feeds is most likely over-estimated because similarity in actions can be homophily.

Vast majority of people's behavior can be predicted by their past actions.





Sharma and Cosley. Distinguishing between Personal Preferences and Social Influence in Online Activity Feeds. CSCW 2015.

PART I. Introduction to Counterfactual Reasoning

PART II. Methods for Causal Inference

PART III. Large-scale and Network Data

PART IV. Broader Landscape