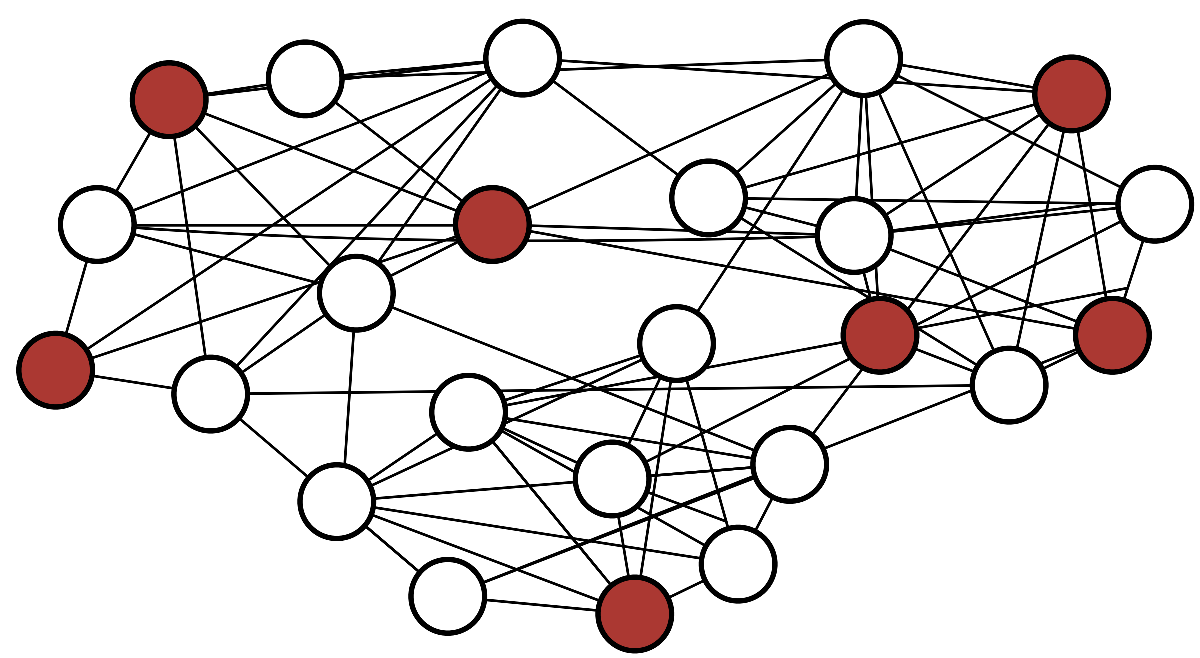
A/B Testing in Networks with Adversarial Members

MLG Workshop 2017

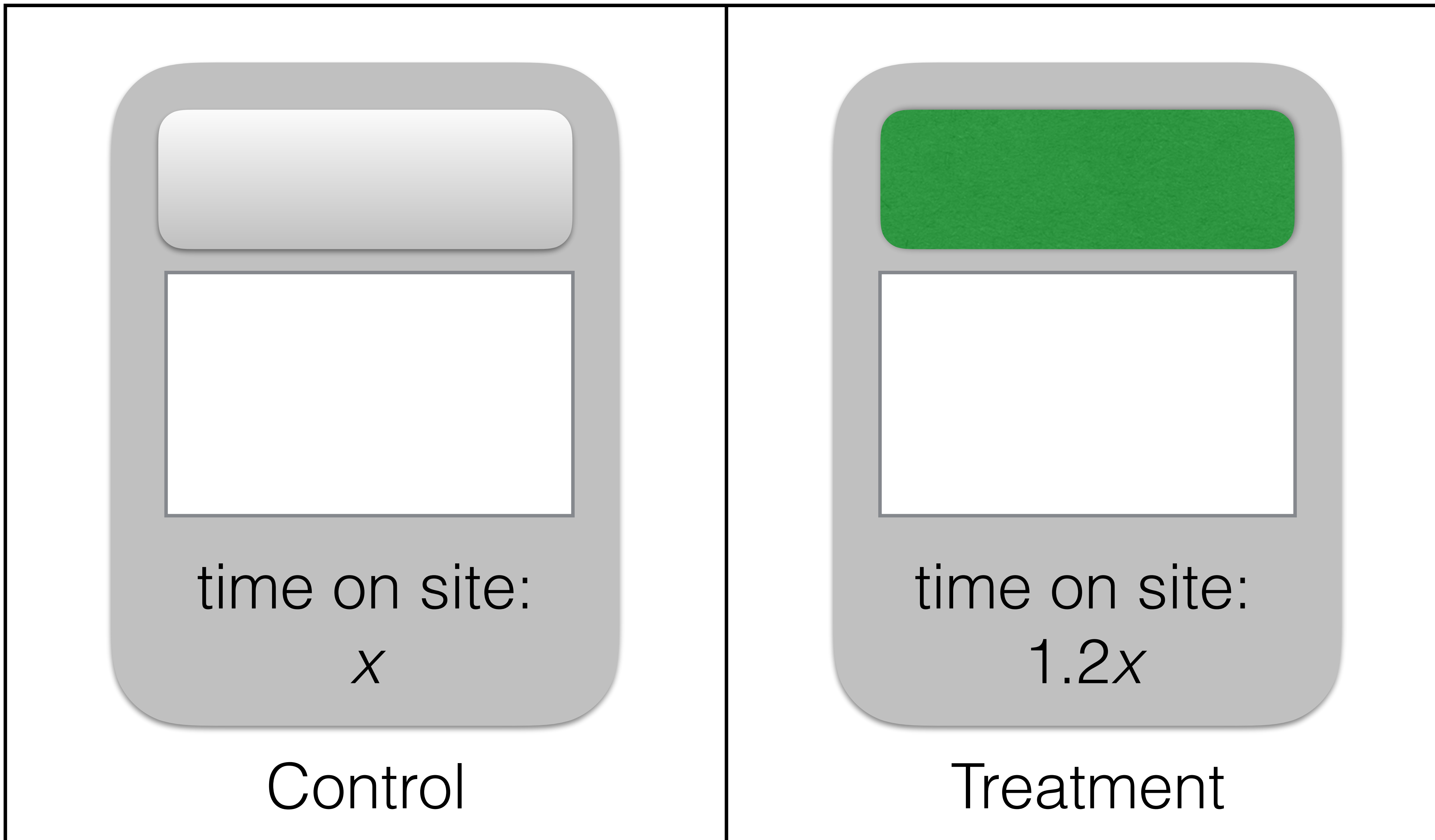
Kaleigh Clary and David Jensen

University of Massachusetts Amherst

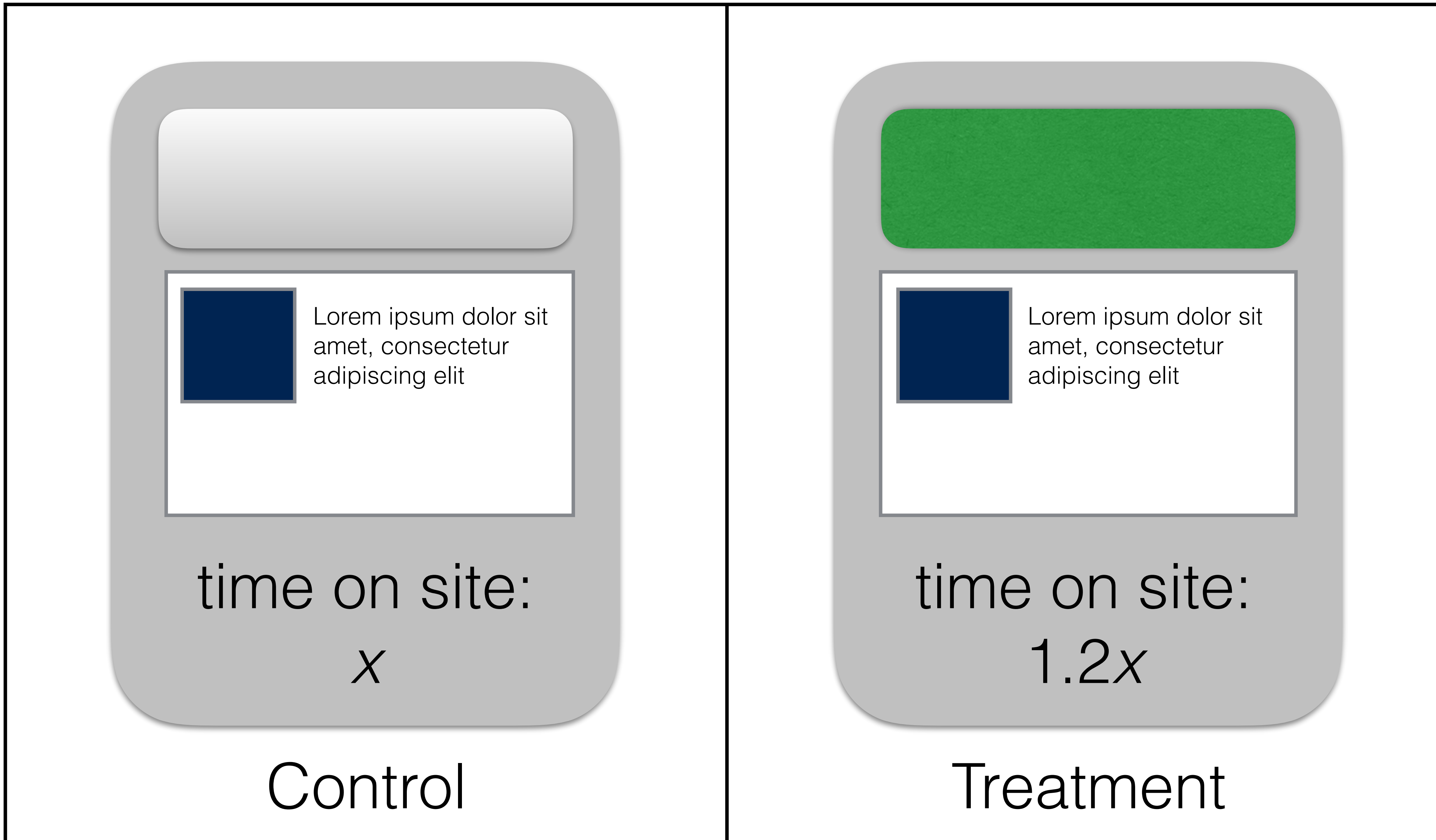
August 14, 2017



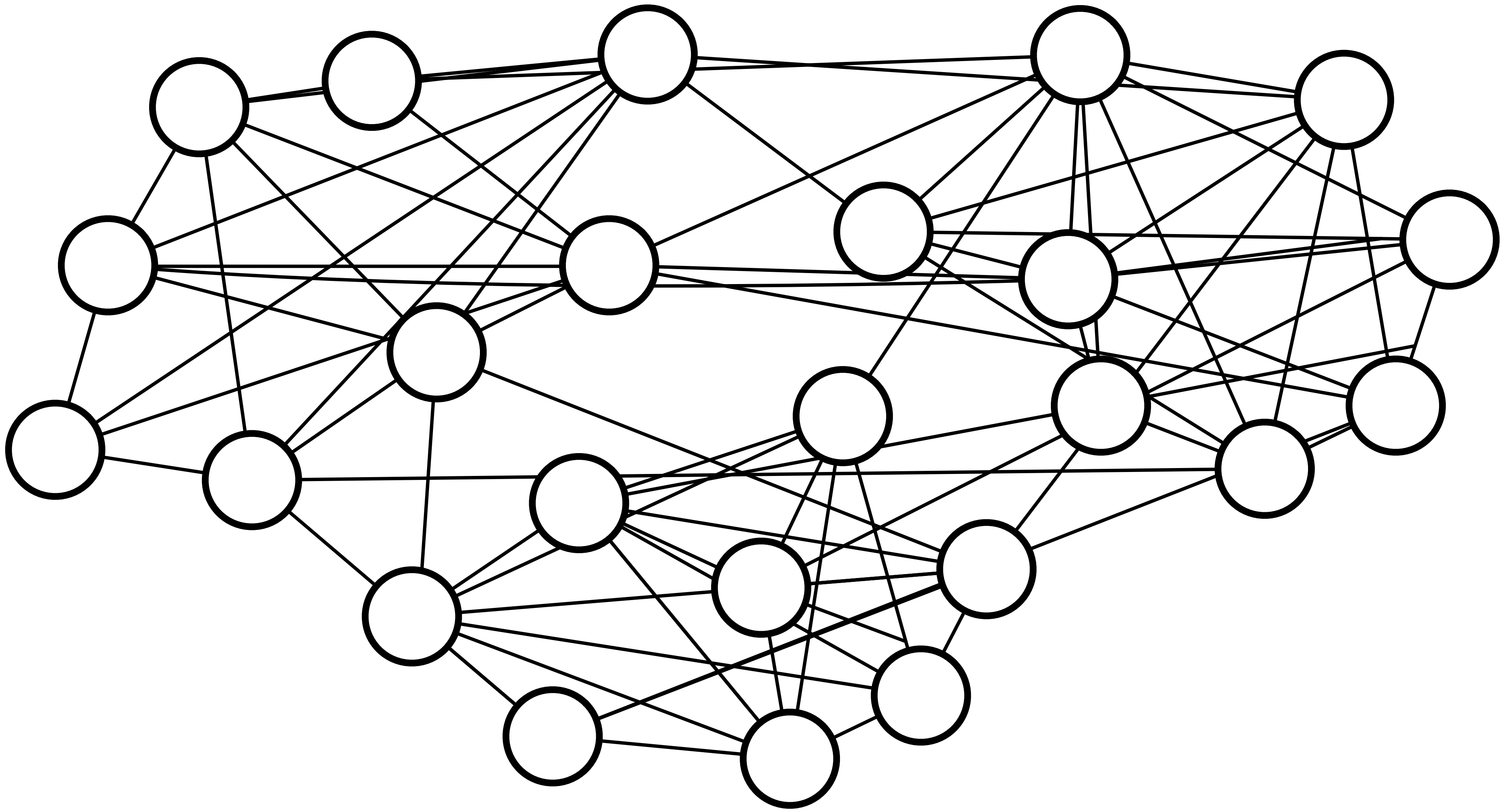
A/B Testing

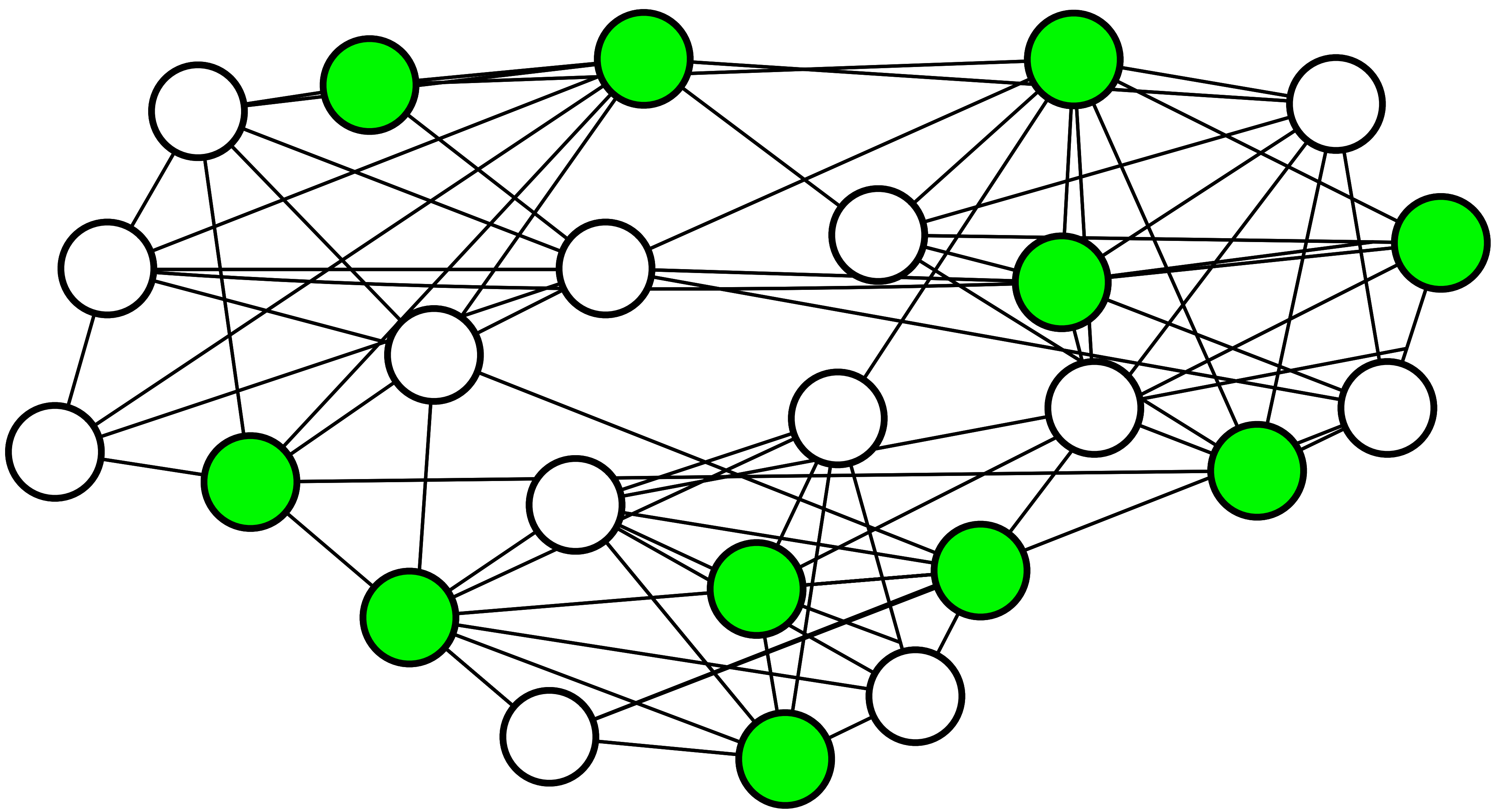


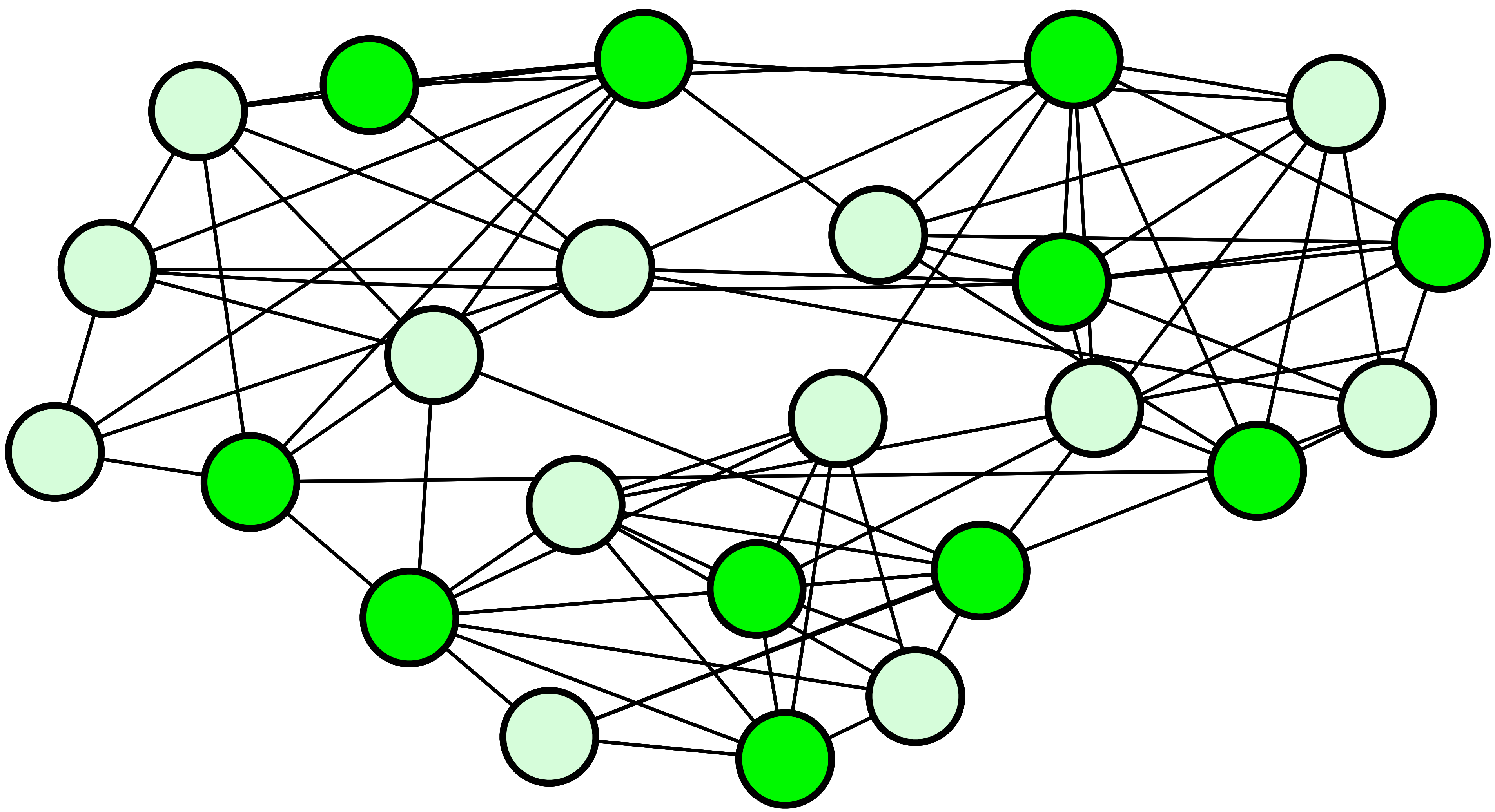
A/B Testing

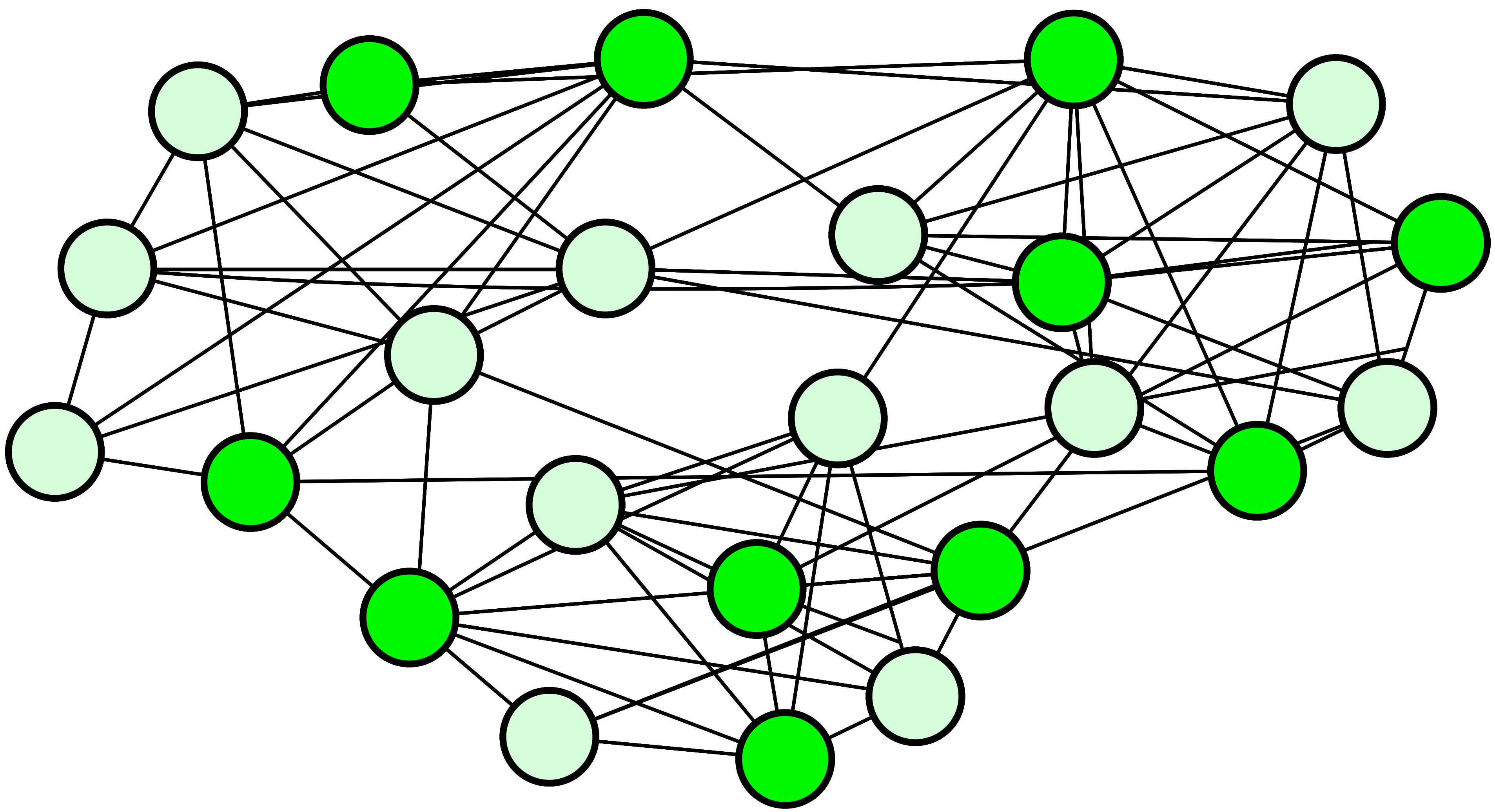


Lorem ipsum dolor sit amet, consectetur adipiscing elit
time on site:
x
Control
Lorem ipsum dolor sit amet, consectetur adipiscing elit
time on site:
1.2x
Treatment



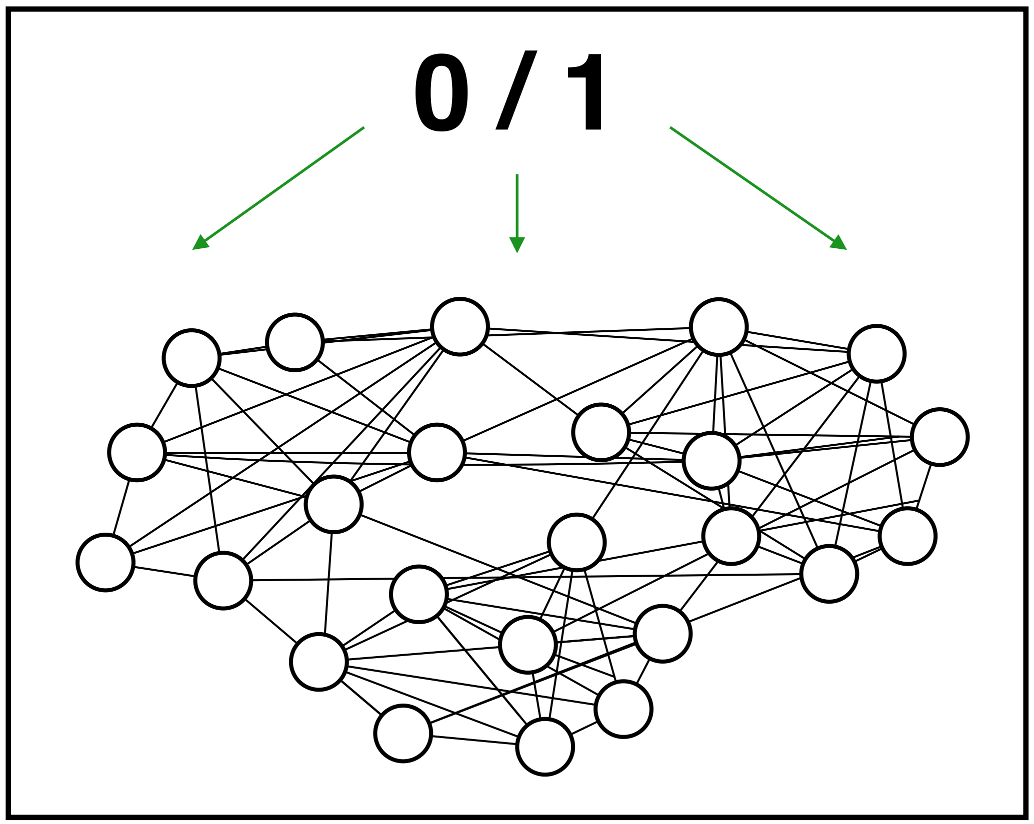


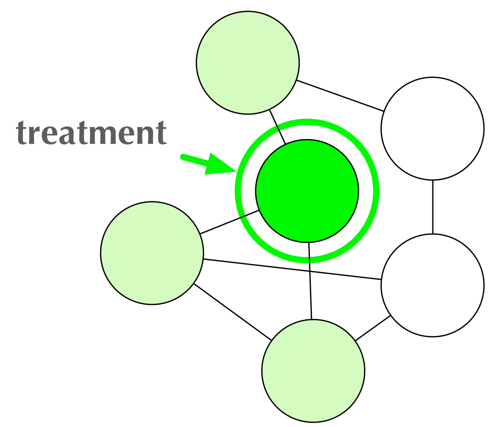




violation of SUTVA

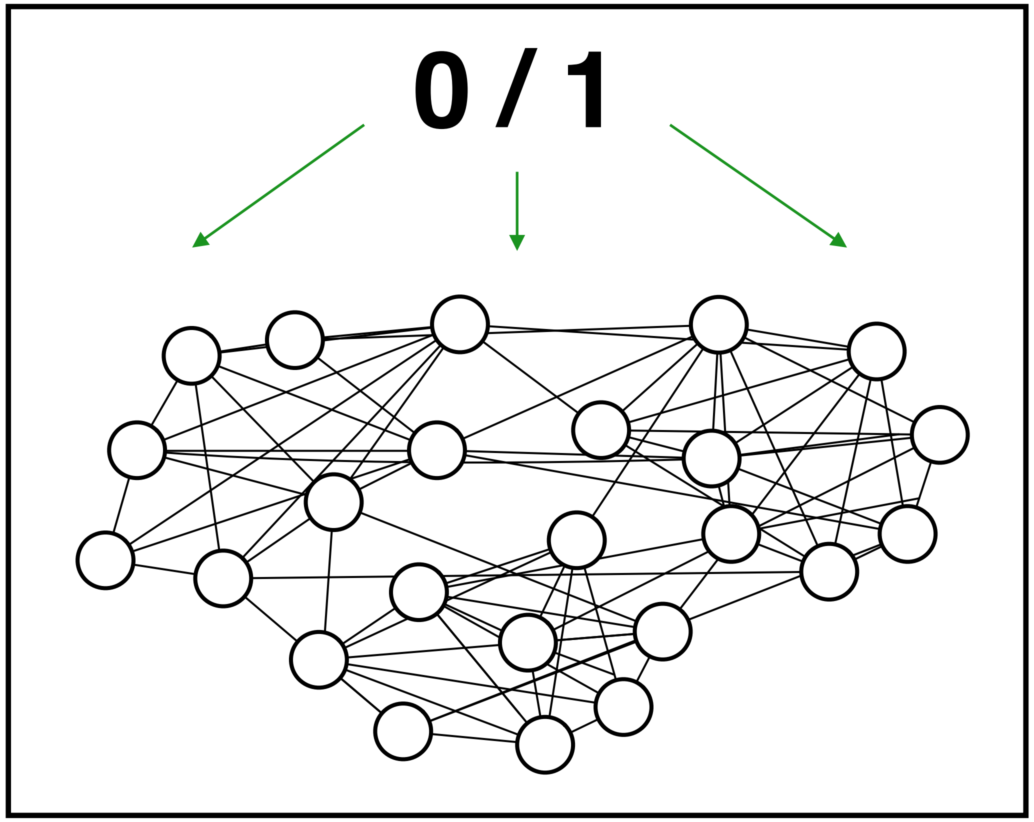
A/B Testing in Networks: Graph Cluster Randomization

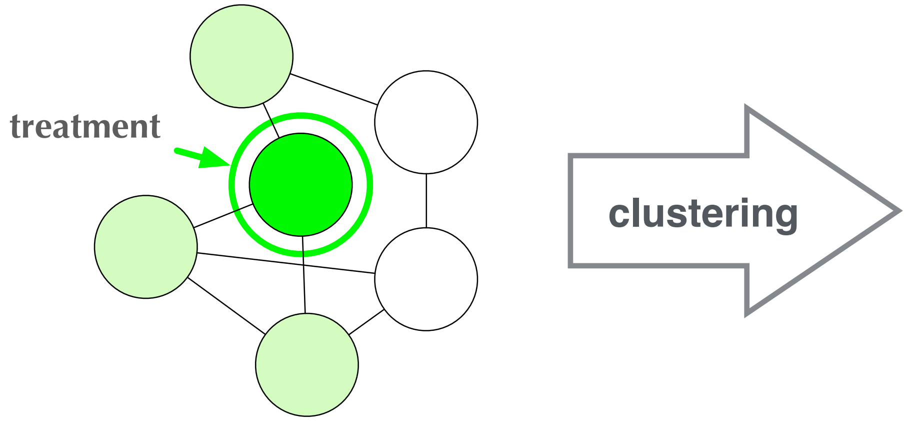




J. Ugander, B. Karrer, L. Backstrom, and J. Kleinberg. Graph cluster randomization: Network exposure to multiple universes. SIGKDD 2013

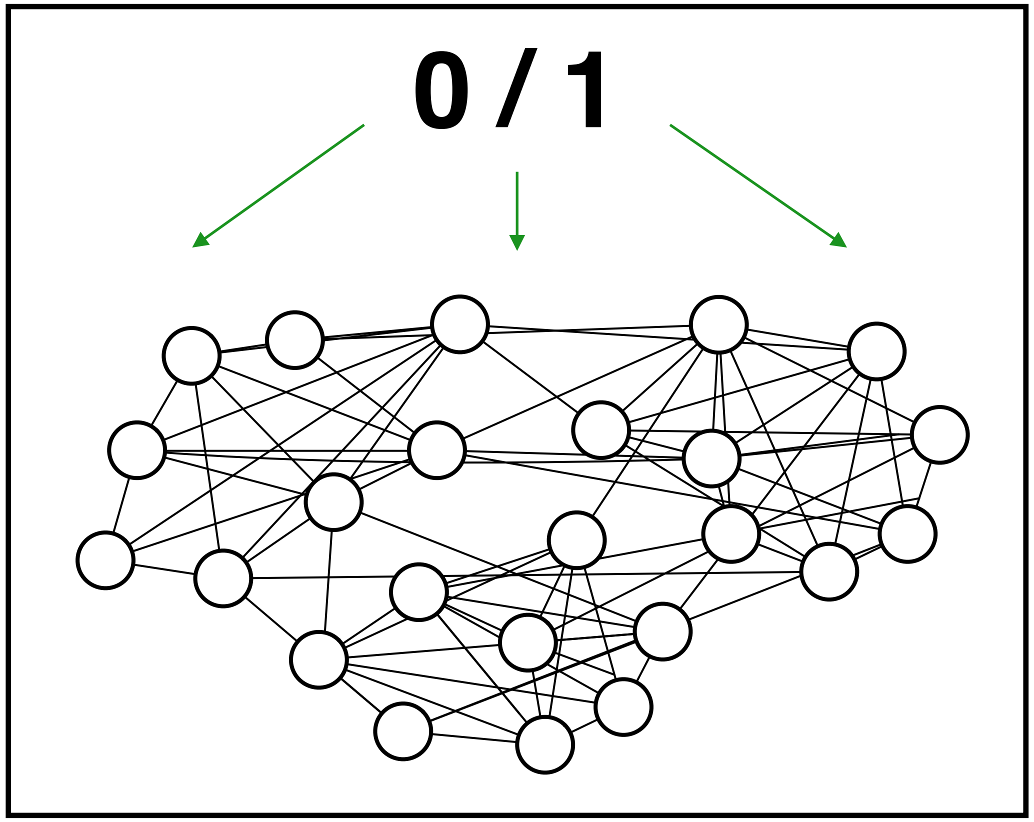
A/B Testing in Networks: Graph Cluster Randomization

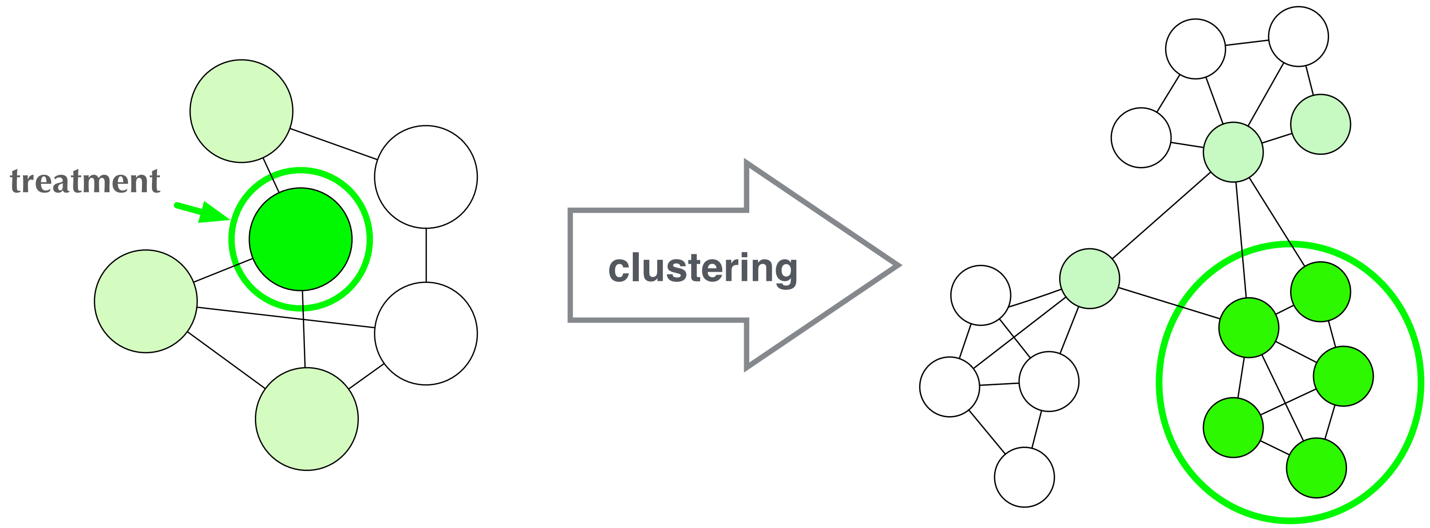




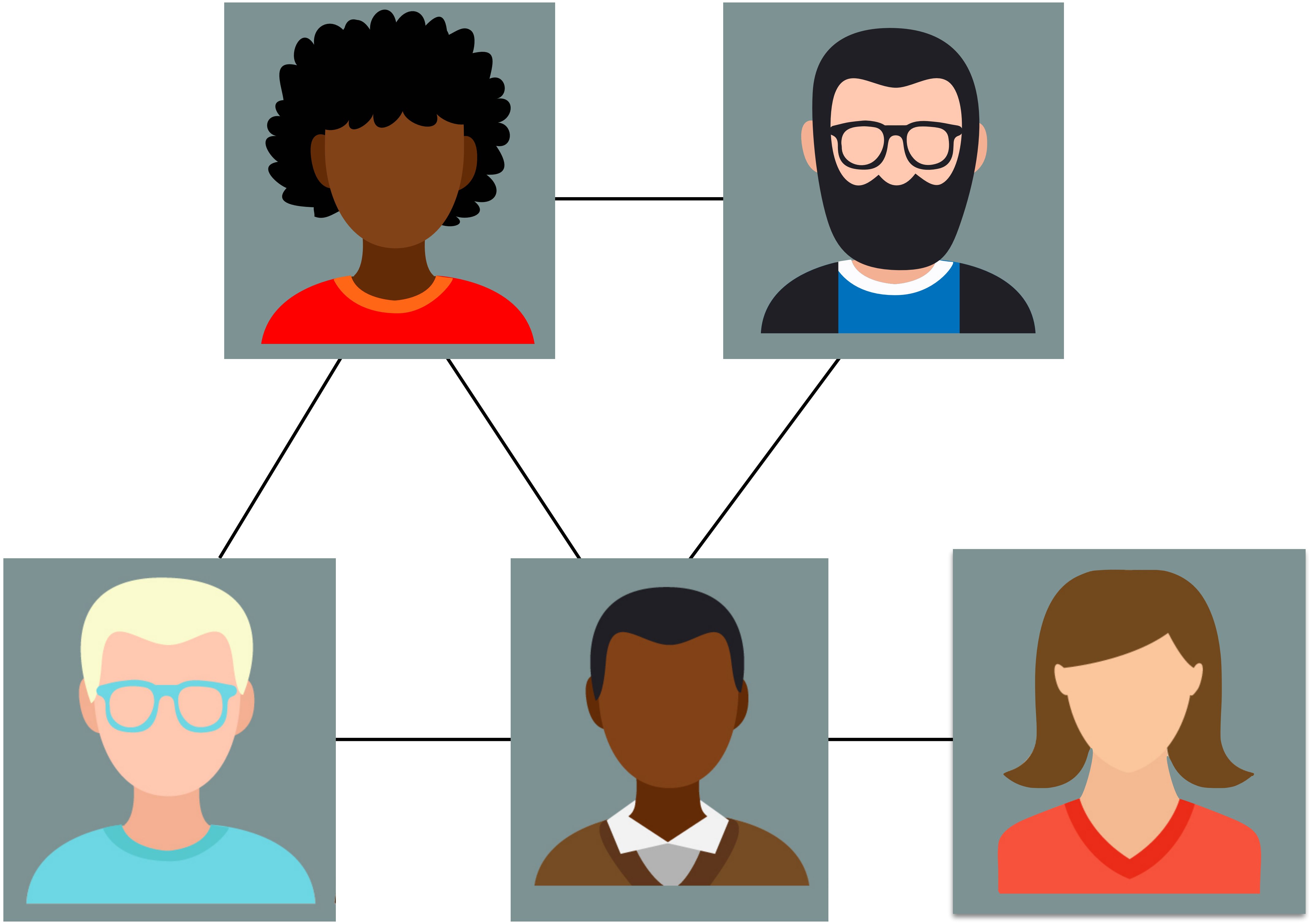
J. Ugander, B. Karrer, L. Backstrom, and J. Kleinberg. Graph cluster randomization: Network exposure to multiple universes. SIGKDD 2013

A/B Testing in Networks: Graph Cluster Randomization

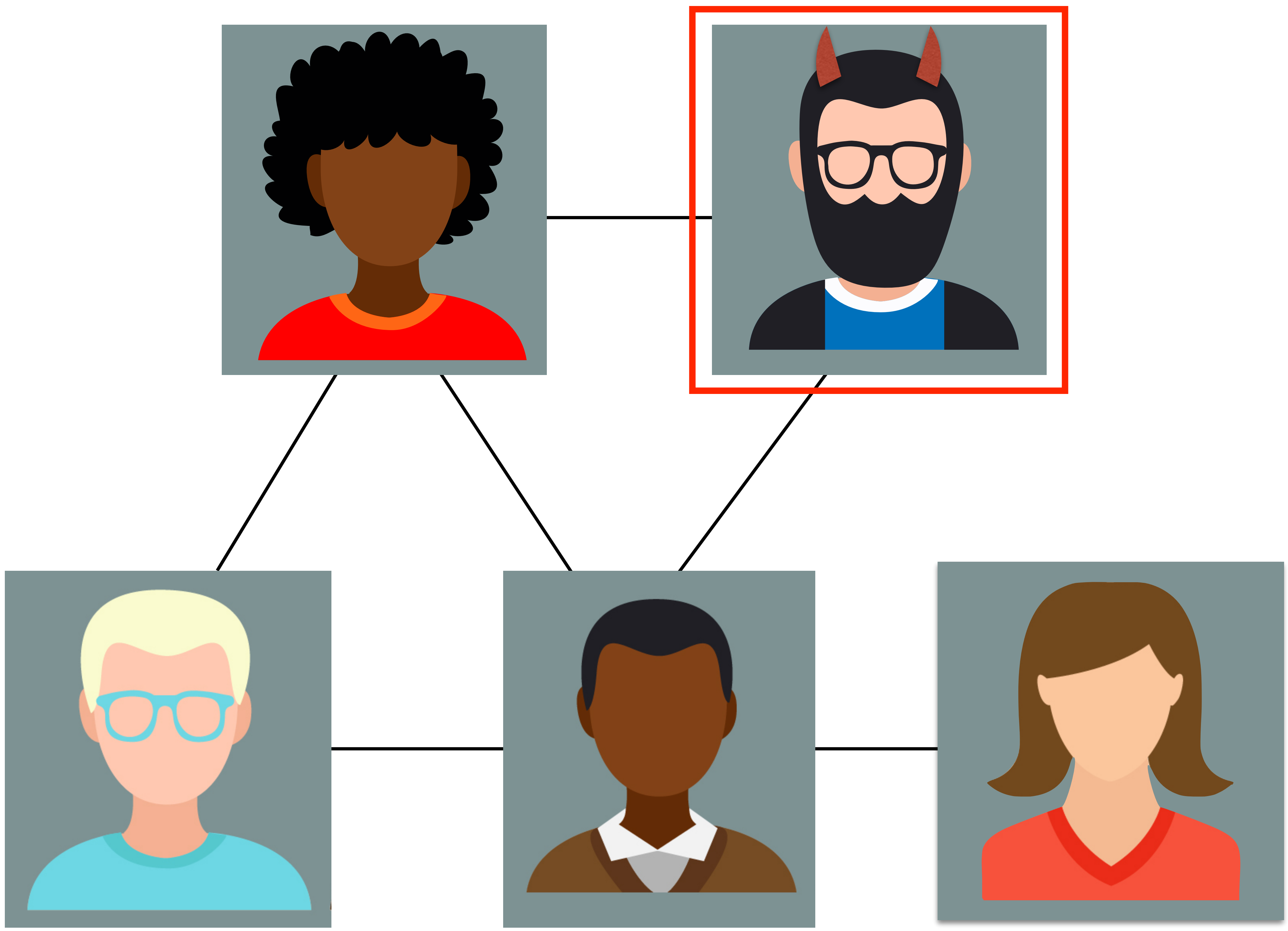




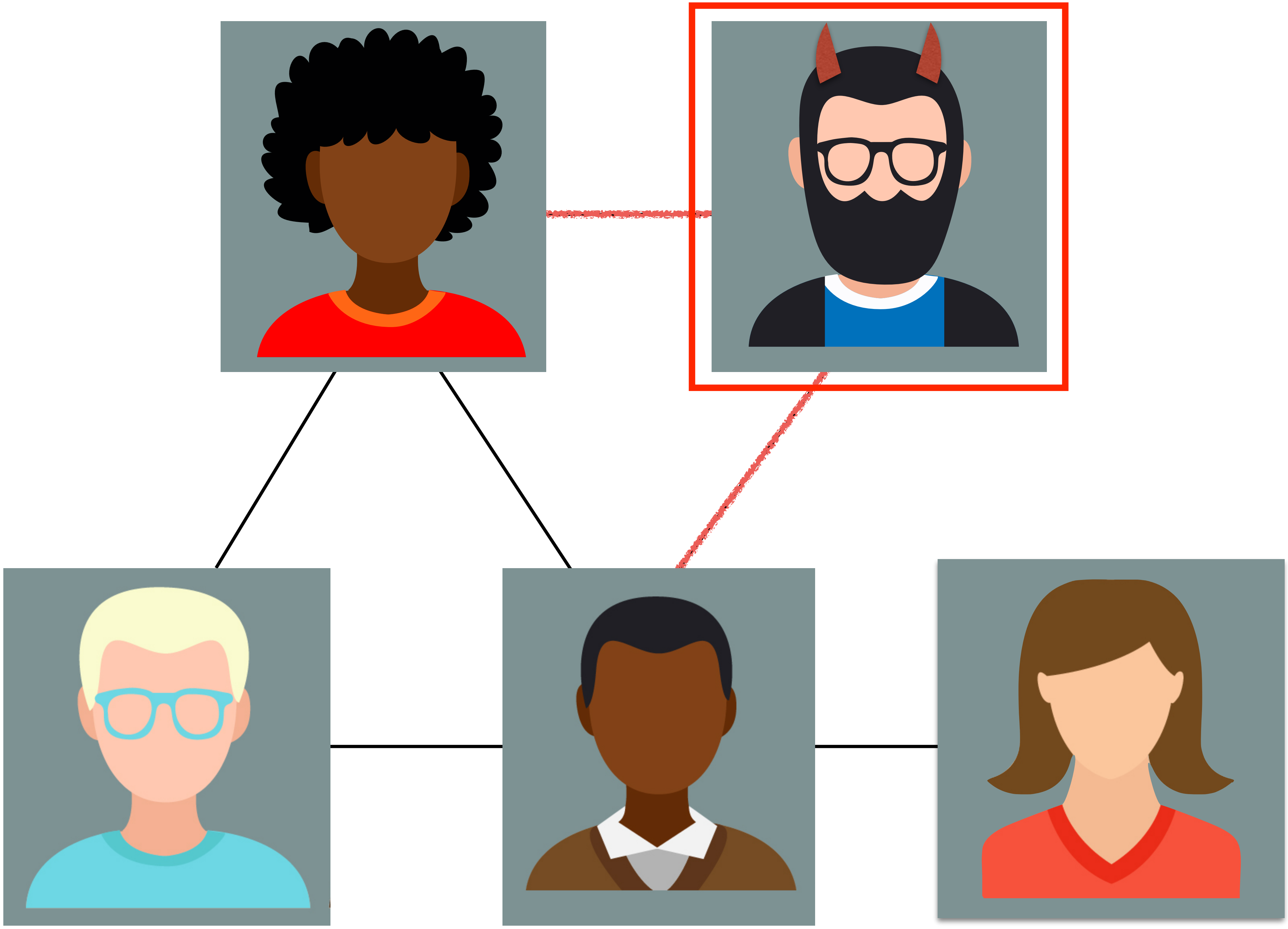
J. Ugander, B. Karrer, L. Backstrom, and J. Kleinberg. Graph cluster randomization: Network exposure to multiple universes. SIGKDD 2013



Avatar vector images designed by Freepik



Avatar vector images designed by Freepik



Avatar vector images designed by Freepik

Adversaries

- Participants in the experiment who would like to influence the estimate of interest, e.g. ATE
    - Increasing or decreasing the estimate of interest
    - Introducing random behavior to the estimate

Adversaries

- Participants in the experiment who would like to influence the estimate of interest, e.g. ATE
    - Increasing or decreasing the estimate of interest → Introducing bias
    - Introducing random behavior to the estimate

Adversaries

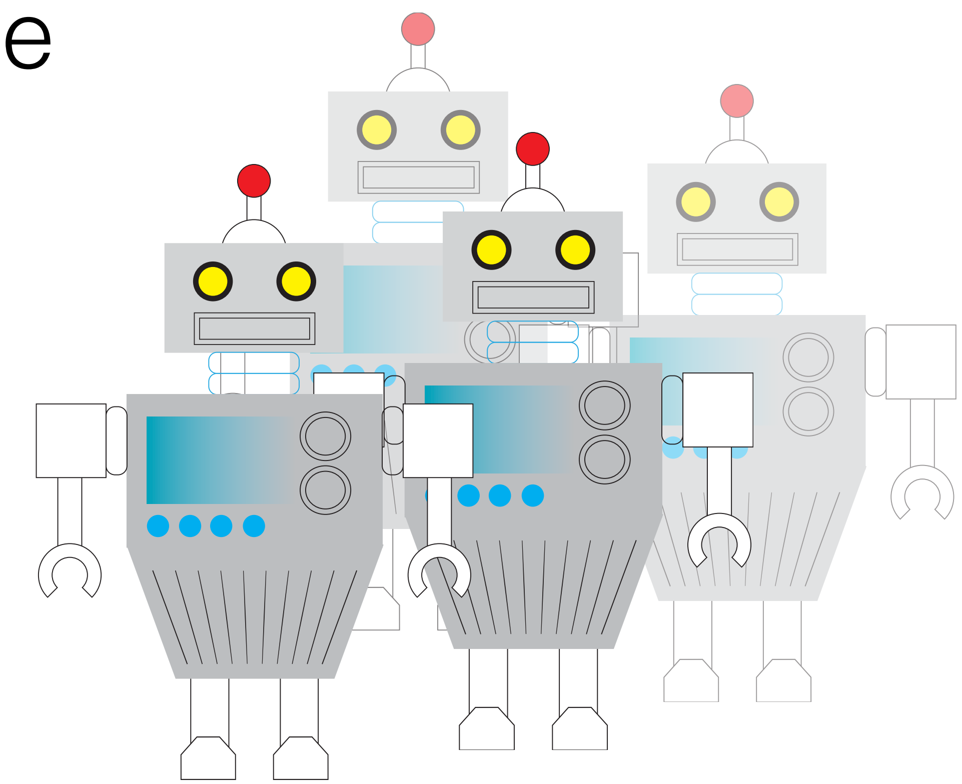
- Participants in the experiment who would like to influence the estimate of interest, e.g. ATE

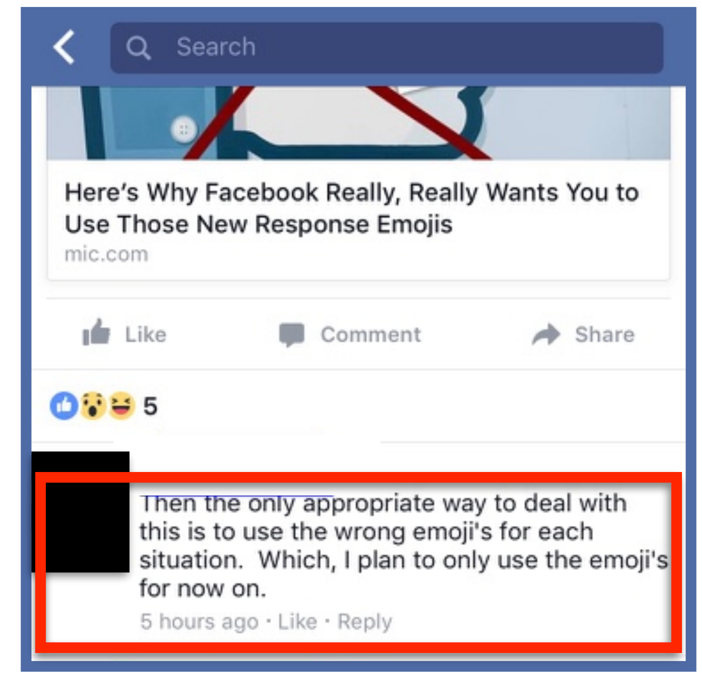
  - Increasing or decreasing the estimate of interest → Introducing bias

  - Introducing random behavior to the estimate → Increasing variance

Motivations for Adversaries

- Competition
- Noncompliance
- Privacy

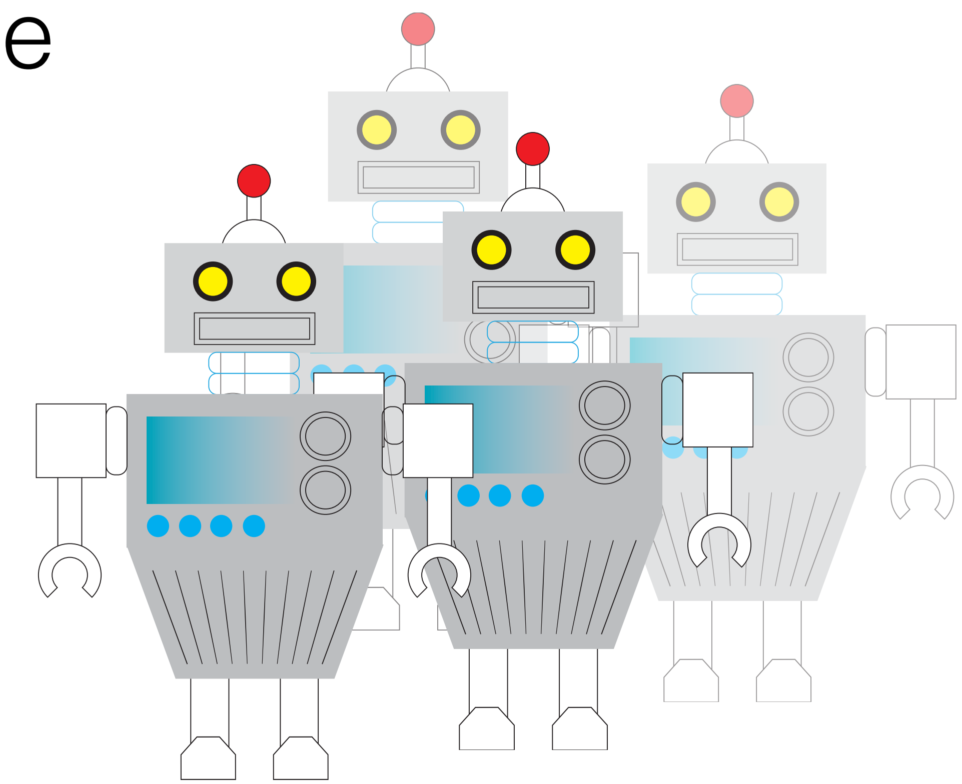


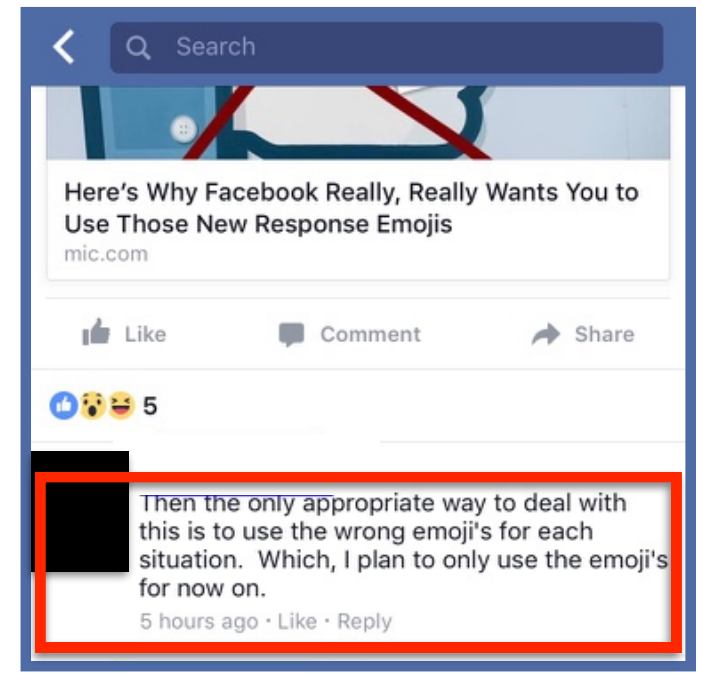


Avatar vector image designed by Freepik
Robot vector images designed by Vecteezy

Motivations for Adversaries

- Competition
- Noncompliance
- Privacy





Assume all adversaries have the same behavior model

Avatar vector image designed by Freepik
Robot vector images designed by Vecteezy

ATE Estimation in Networks

- Assume outcome is a linear additive function of treatment (Gui et al, 2015)

$$Y_i(Z) = \alpha + \beta z_i + \gamma A_i^T Z + \eta A_i^T Y / D_{ii}$$

β: individual treatment effect
γ: peer treatment effect
η: peer outcome effect

- Estimate treatment effect from data

$$g(z_i, \sigma_i) = \alpha + \beta z_i + \gamma \sigma_i$$

z_i: treatment assignment
σ_i: treatment exposure

$$A\hat{T}E = \hat{\beta} + \hat{\gamma}$$

H. Gui, Y. Xu, A. Bhasin, and J. Han. Network A/B testing: From sampling to estimation. WWW 2015

ATE Estimation in Networks

- Assume outcome is a linear additive function of treatment (Gui et al, 2015)

$$Y_i(Z) = \alpha + \beta z_i + \gamma A_i^T Z + \eta A_i^T Y / D_{ii}$$

β: individual treatment effect
γ: peer treatment effect
η: peer outcome effect

- Estimate treatment effect from data

$$g(z_i, \sigma_i) = \alpha + \beta z_i + \gamma \sigma_i$$

z_i: treatment assignment
σ_i: treatment exposure

$$A\hat{T}E = \hat{\beta} + \hat{\gamma}$$

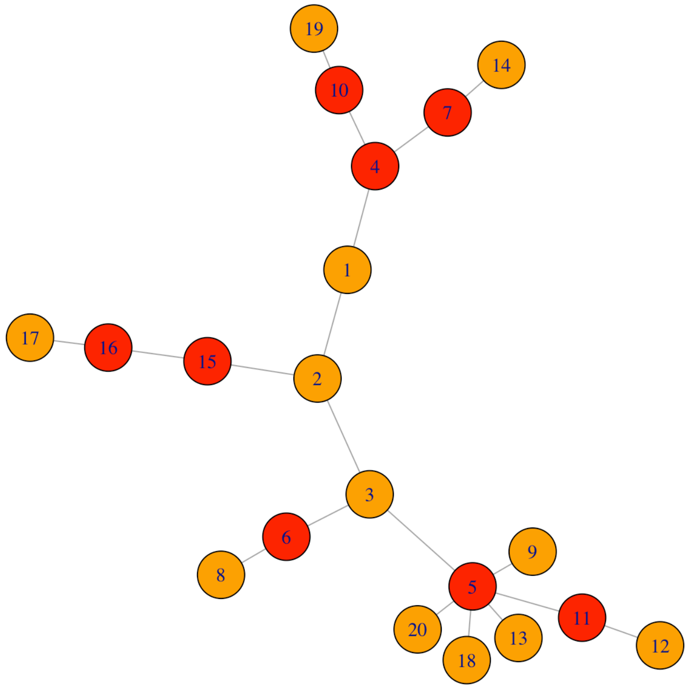
H. Gui, Y. Xu, A. Bhasin, and J. Han. Network A/B testing: From sampling to estimation. WWW 2015

Adversary Placement

(1) Random assignment over the graph

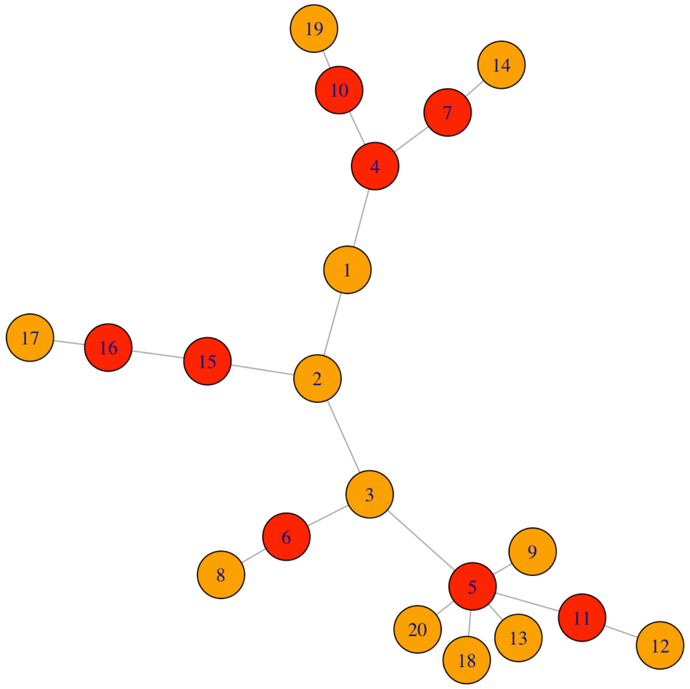
(2) Targeted adversary placement

- When adversaries form a **dominating set** over the graph, every vertex contains at least one adversary in their set of neighbors



Adversary Influence

- Adversary **influence** measures the sum of relative effects of a node on its neighbors' outcomes

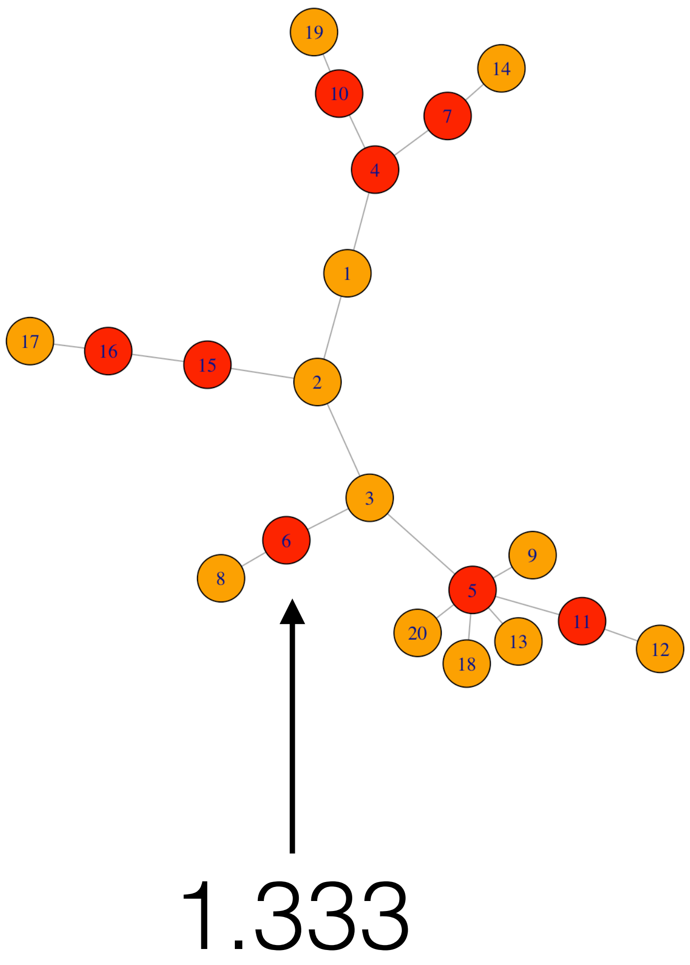


$$\omega_i = D^{-1} A \mathbb{I}_i^N$$

column sum of transition probabilities

Adversary Influence

- Adversary **influence** measures the sum of relative effects of a node on its neighbors' outcomes



$$\omega_i = D^{-1} A \mathbb{I}_i^N$$

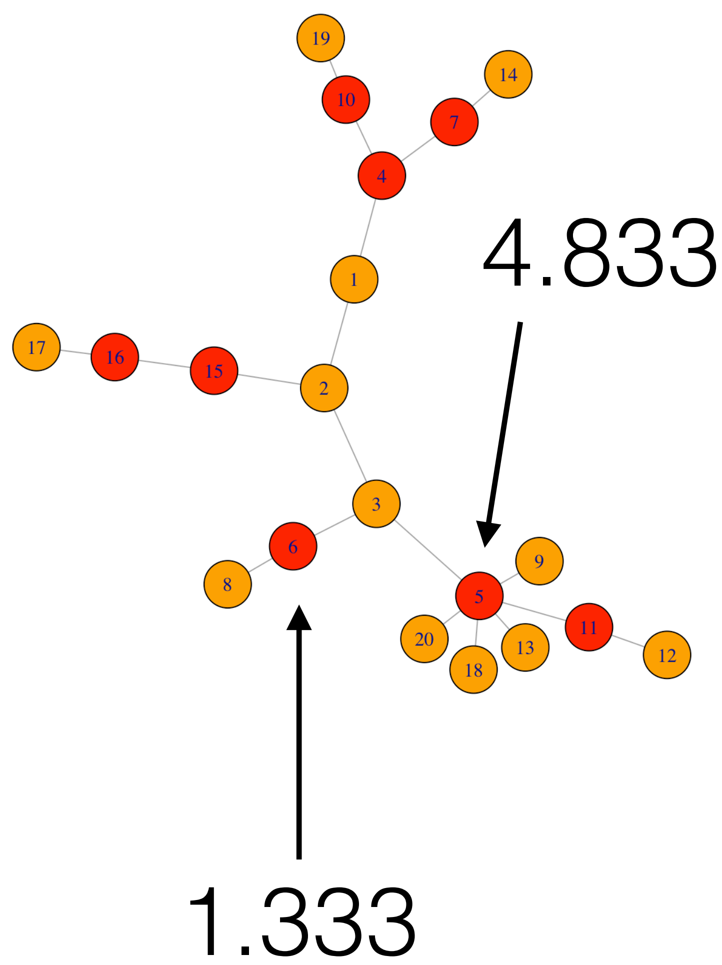
column sum of transition probabilities

Adversary Influence

- Adversary **influence** measures the sum of relative effects of a node on its neighbors' outcomes

$$\omega_i = D^{-1} A \mathbb{I}_i^N$$

column sum of transition probabilities



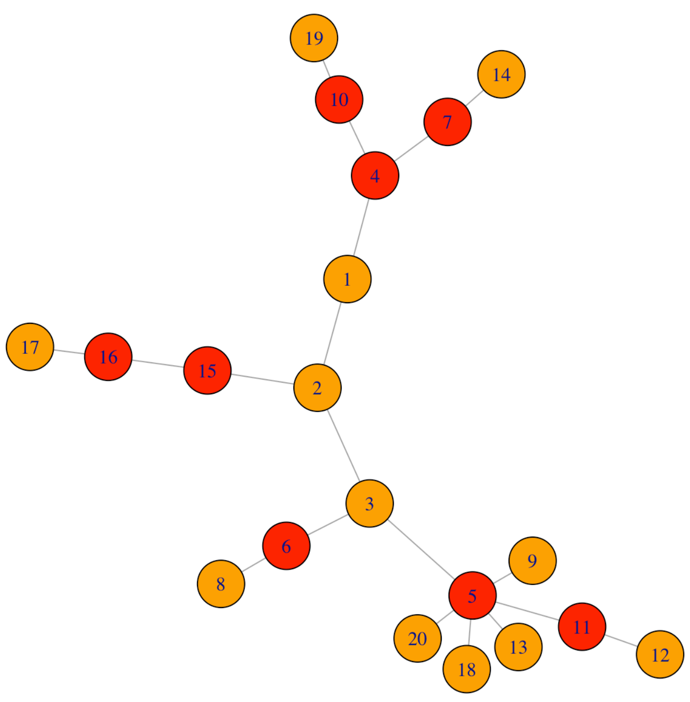
Bias in ATE Induced by Adversaries

- Adversaries bias ATE estimates through:

  (1) The value of their outcome

  (2) The effect of their outcome on their neighbors' outcome

  - strength depends on true network effect



$$\hat{ATE}_{R_Y} = \sum_{j \in A_r} \frac{1}{d_j}(Y_r - \bar{Y}_{A_j \setminus r})$$

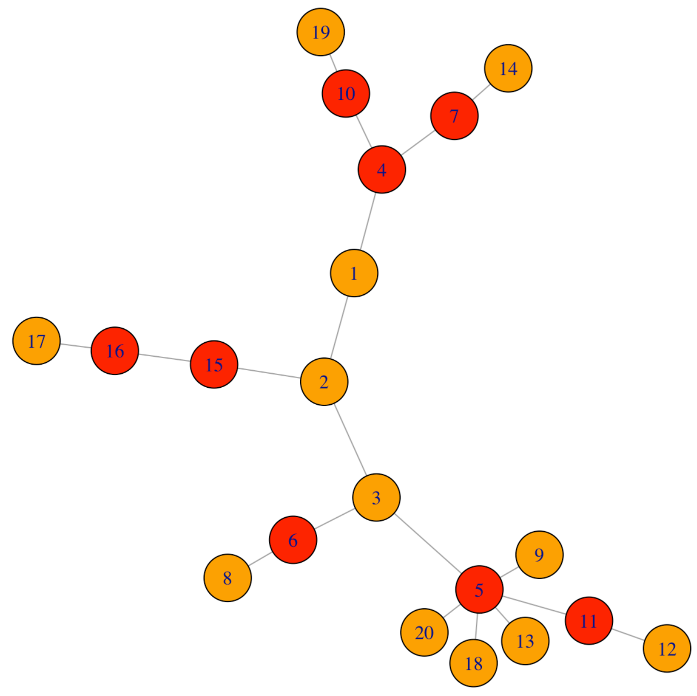
Bias in ATE Induced by Adversaries

- Adversaries bias ATE estimates through:

    (1) The value of their outcome

    (2) The effect of their outcome on their neighbors' outcome

        - strength depends on true network effect



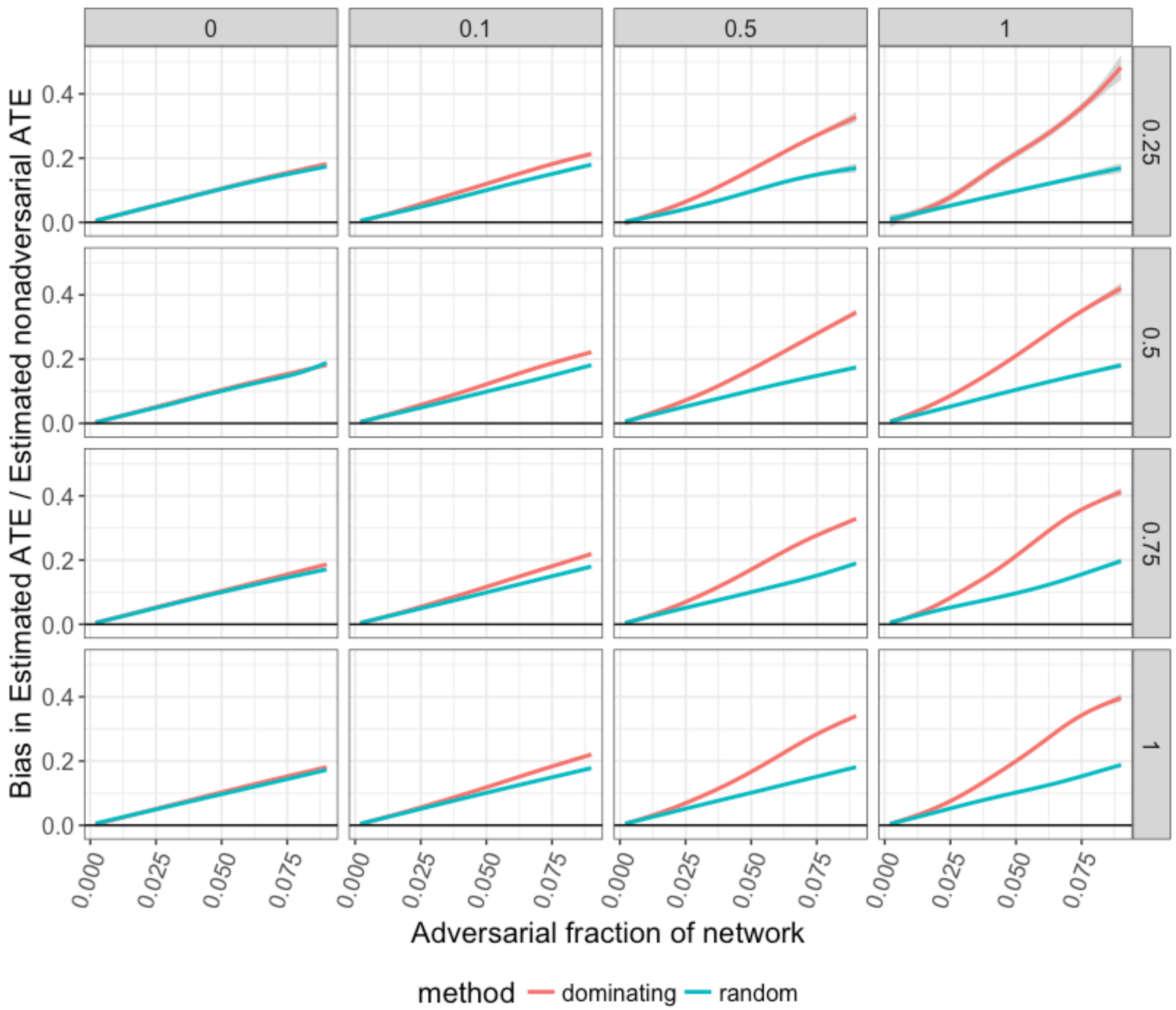
$$\hat{ATE}_{R_Y} = \sum_{j \in A_r} \frac{1}{d_j} (Y_r - \bar{Y}_{A_j \setminus r})$$

$$\approx \omega_r (Y_r - \bar{Y}_{A_{2_r}})$$

Approximate bias from effect of adversary outcome using *influence, ω*

Experimental Results

λ_2 Peer Effect



λ_1

Treatment Effect

Normal

$$Y_{i,t} = \lambda_0 + \lambda_1 z_i + \lambda_2 \frac{A_i Y_{t-1}}{D_{i,i}} + U_{i,t}$$

Adversary

$$Y_r = \begin{cases} \lambda_0 & \text{if } z_r = 1, \\ \lambda_0 + \lambda_1 & \text{if } z_r = 0. \end{cases}$$

Summary

- Derived expressions for the bias induced by adversary behavior
- Empirically demonstrated a vulnerability in network A/B testing to manipulation of ATE estimates from exploitation of peer effects
- Examined the difference between random and targeted placement of adversaries in the network
- *Future work*: Characterize the relationship between adversary detection and strength of adversarial response

Thank you