Optimal Bayesian Designs for Network A/B Tests

Trang Bui
tqtbui@uwaterloo.ca

Department of Statistics and Actuarial Sciences
University of Waterloo

June 19, 2024

Online A/B testing

- We now live in an online world
  - Smartphones
  - Social media
  - Online shopping
  - etc.

- A lot of data is now collected online
  - Surveys
  - Experiments!

Figure 1: Online world[1]

[1] Image retrieved from https://proschoolonline.com/blog/6-types-of-people-you-must-add-to-your-network

A/B testing:

- Experiments with one factor at two levels: control (A) and treatment (B).

- Examples:
  - Old web design vs new web design
  - Old advertisement vs new advertisement

- Should Facebook deploy the new update to everyone?

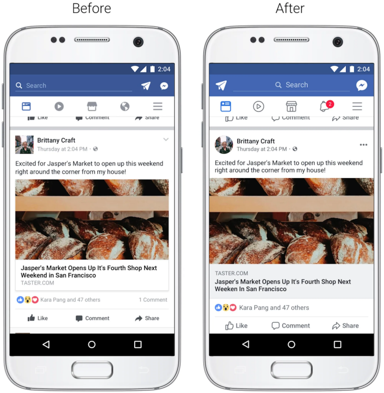
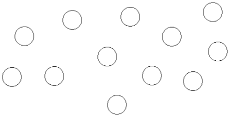


Figure 2: Facebook interface update[2]

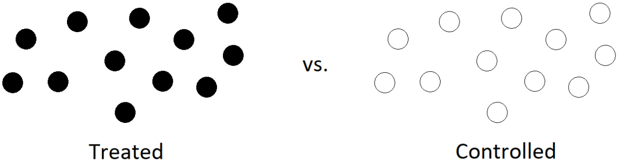
Image retrieved from https://9to5mac.com/2017/08/15/facebook-updates-camera-newsfeed/

Global treatment effect

- Suppose that Facebook has access to n units (users) to answer this question.



- We want to compare what happens when everyone is treated versus when everyone is controlled.



Global treatment effect (cont.)

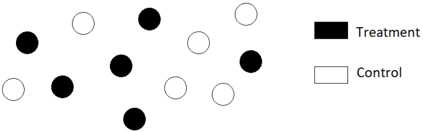
- The Global Average Treatment Effect (GATE) quantifies this.
  - Let Y_i denote the outcome of unit i and $Z_i \in \{0, 1\}$ be a treatment indicator.
  - The GATE is defined as

$$\text{GATE} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i \middle| \boldsymbol{Z} = \mathbf{1}_n\right] - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i \middle| \boldsymbol{Z} = \mathbf{0}_n\right]$$

- We cannot observe both everyone being treated and everyone being controlled.
  - GATE is estimated by assigning some units to treatment and some units to control.

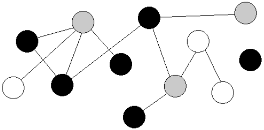
Classical vs Network A/B testing

Classical A/B testing:



- Units are independent

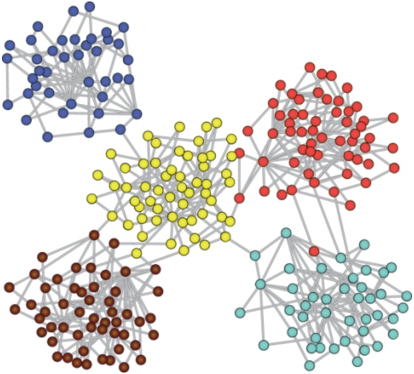
Network A/B testing



- Units correlate and interfere with one another

Design-based approach

- Graph cluster randomization [1], [2]
  - Cluster graphs
  - Assign each cluster to treatment or control
- Goal: Estimate GATE
  - Compare average outcomes from treated and controlled clusters
- Issues
  - Clustering graphs is hard!
  - Discard borderline units.
  - Only focus on GATE



Model-based approach

- Notation
  - Experimental units: $1, 2, ..., n$
  - Y_i: outcome of unit i
  - $Z_i \in \{0, 1\}$: be a treatment indicator.
- Assume a generating model for the experimental outcomes:

$$\mathbf{Y} \sim p_{\mathbf{Y}}(\mathbf{D}, \boldsymbol{\theta});$$

where $\mathbf{D} = \{\mathcal{G}, \mathbf{Y}, \mathbf{Z}, \mathbf{X}\}$ denote experimental data.

  - $\mathcal{G}$: the network (graph)
  - $\mathbf{X}$: possible covariates
  - Analysis conditional on the given treatment assignment vector.
  - Utilize all units for analysis, provide more possible inferences.

Example models

- Power-Degree (POW-DEG) model [3]:

$$Y_i = \mu + \tau Z_i + \gamma_1 \left(\sum_{j \in N(i)} Z_j \right)^{\lambda} + \gamma_2 \left(\sum_{j \in N(i)} (1 - Z_j) \right)^{\lambda} + \epsilon_i$$

- Conditional Network Autoregressive (CNAR) model [4]:

$$Y_i = \mu + \tau Z_i + \epsilon_i$$

$$\epsilon_i | \boldsymbol{\epsilon}_{-i} \sim \mathcal{N} \left(\rho \sum_{j \in N(i)} \frac{\epsilon_j}{k_i}, \frac{\sigma^2}{k_i} \right)$$

Example models (cont.)

- Normal Sum (NS) model [5]:

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$
$$Y_i(0) \mid \boldsymbol{X} \sim \mathcal{N}\Big(X_i + \sum_{j \in N(i)} X_j, \gamma^2\Big)$$
$$Y_i = Y_i(0) + \tau Z_i$$

- Binary Network-Temporal Autoregressive (BNTAR) model [6]:

$$Y^*_{i,t} = \mu + \tau Z_i + \gamma \frac{1}{k_i} \sum_{j \in N(i)}^{n} Y_{j,t-1} + \epsilon_{i,t}$$
$$Y_{i,t} = \mathbb{I}(Y^*_{i,t} > 0)$$

Model-based optimal designs

- Design criterion:

$$\phi(\mathbf{Z}) = L_{p_{\mathbf{Y}}}(\mathbf{D}, \boldsymbol{\theta}) = \text{MSE(GATE)}$$

- Optimal design:

$$\mathbf{Z}^* = \underset{\mathbf{Z} \in \{0,1\}^n}{\arg\min} \; L_{p_{\mathbf{Y}}}(\mathbf{D}, \boldsymbol{\theta})$$

- Notes: Units are non-exchangeable, meaning the design is deterministic.
    - Here, $\mathbf{Z}$ is the design of the experiment.

Challenges

There are two main challenges:

- Evaluate the design criterion $\phi(\mathbf{Z}) = L_{p_Y}(\mathbf{Z}, \boldsymbol{\theta})$
  - Outcome-generating models are complex.
  - Design criterion involves unknown parameters.
- Construct the optimal design $\mathbf{Z}^* = \underset{\mathbf{Z}\in\{0,1\}^n}{\arg\min}\, L_{p_Y}(\mathbf{Z}, \boldsymbol{\theta})$
  - Discrete and exponentially large search space: $\mathbf{Z} \in \{0, 1\}^n$.
  - Problem is NP-hard!

Bayesian Optimal Design Criterion

- We propose to use a Bayesian design criterion

$$\phi(\mathbf{Z}) = \int_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L_{p_Y}(\mathbf{Z}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

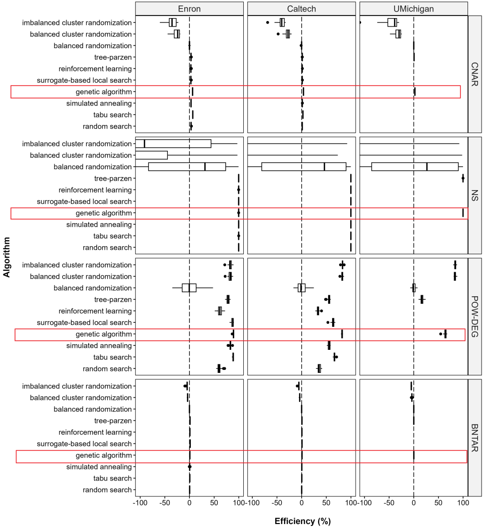
  - Enhance the robustness of the design to parameter misspecification;
  - However, ϕ does not have a closed-form formula
    - Monte Carlo approximation.

Bayesian Optimal Design Construction

- Adapt algorithms to construct optimal designs on the discrete space
  - Meta-heuristics: Random Search, Tabu Search, Simulated annealing, Genetic algorithms.
  - Bayesian optimization: NN surrogates, REINFORCE algorithm, Tree-Parzen estimator.
  - Can be applied to any model and design criterion.
- Simulations
  - Focus on MSE(GATE)
  - Construct (sub)-optimal designs for various networks and data-generating models.
  - Compare to graph-cluster randomization and naive randomization.

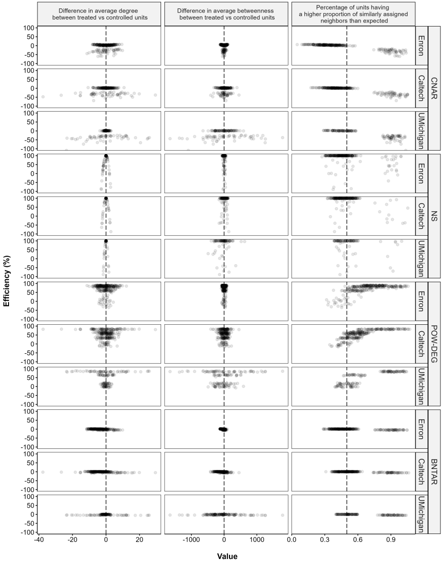
Simulation Results: Comparison of Algorithms

- Genetic algorithm is easy to train, fast, and gives good results.
- Graph cluster randomization is not a universally good design for the GATE.



Simulation Results: Characteristics of Good Designs

- Plot characteristics of (sub)-optimal designs found by different algorithms.
- There is no one-size-fits-all solution.
- Good designs tend to be balanced in terms of allocations and network characteristics.



Conclusions and Future Directions

- Contributions:
  - ✓ Enhanced robustness of the model-based approach
  - ✓ Construction of optimal designs accounting for parameter uncertainty
  - ✓ Adapt design algorithms to apply to any model and design criterion
- Future directions:
  - Randomized designs.
  - Model-agnostic designs.
  - Designs over time.

Bibliography

[1] H. Gui, Y. Xu, A. Bhasin, and J. Han, "Network a/b testing: From sampling to estimation," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 399–409.

[2] J. Ugander, B. Karrer, L. Backstrom, and J. Kleinberg, "Graph cluster randomization: Network exposure to multiple universes," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 329–337.

[3] T. Bui, S. H. Steiner, and N. T. Stevens, "General additive network effect models," *The New England Journal of Statistics in Data Science*, pp. 1–19, 2023, ISSN: 2693-7166. DOI: 10.51387/23-NEJSDS29.

[4] Q. Zhang and L. Kang, "Locally optimal design for A/B tests in the presence of covariates and network dependence," *Technometrics*, vol. 64, no. 3, pp. 358–369, 2022.

[5] G. W. Basse and E. M. Airoldi, "Model-assisted design of experiments in the presence of network-correlated outcomes," *Biometrika*, vol. 105, no. 4, pp. 849–858, 2018.

[6] D. Eckles, B. Karrer, and J. Ugander, "Design and analysis of experiments in networks: Reducing bias from interference," *Journal of Causal Inference*, vol. 5, no. 1, 2016.

Thank you!