Beyond the Experiment Window: Prospective Impacts Under Long-Term Ranking Dynamics

Lei Shi, UC Berkeley | Lo-Hua Yuan, Airbnb | Peng Ding, UC Berkeley | Navin Sivanandam, Airbnb

Motivation: The Peril of Short-Term Optimization

What works in the short run does not always work in the long run. Imagine a search ranking algorithm (ranker) that favors cheaper, low-quality products: short-term conversion might spike, only to be followed by a drop in repeat purchases as customer trust erodes. This is a core challenge of using short-duration A/B tests to evaluate rankers. Operational processes and constraints require successive decision-making on short timescales. Yet these online experiments, often lasting only days or weeks, fail to capture critical longitudinal dynamics—such as seasonality, user evolution, and feedback loops between ranker outputs and the search ecosystem—that dictate long-term platform health. Relying solely on short-term metrics can lead to myopic decisions that optimize for immediate gains at the expense of future growth.

Causal Estimand: Prospective Long-Term Average Treatment Effect

Our objective is to align iterative ranker deployment decisions with strategic long-term goals. Suppose we develop a new ranker and run a month-long A/B test to compare the new and old rankers. At the end of the experiment, we need to decide, "Should we launch the new ranker to all users or revert everyone back to the old ranker?" To formalize this question, consider longitudinal data where observations are in a time sequence order:

$$X_0 \to Z_1 \to S_1 \to Y_1 \to X_1 \to Z_2 \to S_2 \to Y_2 \to \cdots \to X_{T-1} \to Z_T \to S_T \to Y_T.$$

$(\overline{X}_{t-1}, \overline{Z}_t, \overline{S}_t, \overline{Y}_t)$ is a tuple of user covariates (which can be expanded to include item characteristics, macroeconomic conditions, etc.), treatments (ranker version), surrogates, and outcomes. Here, $\overline{a}_t = (a_1, \ldots, a_t)$ denotes the sequence of all past values of a_t, and $\underline{a}_t = (a_t, \ldots, a_T)$ denotes the sequence of all future values of a_t. $Y_t(\overline{Z}_T)$ is the potential outcome at time point t for a user who receives treatment $\overline{Z}_T = \overline{z}_T$. Under the *no anticipation* assumption that potential outcomes at time t are not affected by the future, $Y_t(\overline{z}_T) = Y_t(\overline{z}_t)$. Moreover, under the assumption of *consistency*, the observed outcome Y_t is equal to the potential outcome $Y_t(\overline{Z}_t)$. Formally, our causal estimand of interest is the **prospective long-term average treatment effect (PLATE)**:

$$\tau = \sum_{t=1}^{T} \mathbb{E}\left\{ Y_t(\overline{\zeta}_t{}^{+}) - Y_t(\overline{\zeta}_t{}^{0}) \mid D = e \right\}. \tag{1}$$

This is the average causal effect that a change from old ranker ζ^0 to new ranker ζ^+ has on long-term cumulative outcomes, averaged over all users in the experimental dataset $\{D = e\}$. We focus on the long-term outcome of a *long-lasting intervention* [6]. This distinguishes our estimand from causal quantities studied in the dynamic treatment regimes [4] and long-term impact attribution literature [2], which focus on the long-term outcome of a short-term intervention.

Two data sources. The experimental sample for which we want to estimate PLATE consists of N_e users with observations over $T_{\text{SHORT}} < T$ time points. We also have access to a (historical) observational dataset $\{D = o\}$ consisting of N_o users over $T_{\text{LONG}} \geq T$ time points. Users in the observational data are exposed to rankers ζ^- and ζ^0, while those in the experiment are exposed to rankers ζ^+ and ζ^0. The treatments (ranker version) satisfy a *sequential randomization* assumption:

$$\text{for } t \geq r, \text{ and } d \in \{o, e\}, \quad Y_t(\overline{z}_{r-1}, \underline{z}_r) \perp\!\!\!\perp Z_r \mid \overline{Z}_{r-1} = \overline{z}_{r-1}, \overline{S}_{r-1}, \overline{X}_{r-1}, D = d. \tag{SR}$$

Key Challenges in Estimation

Estimating PLATE is hard. In practice, ranker experiments are short and sequential, and do not capture salient longitudinal dynamics. This leads to three key estimation challenges:

1. **Identifying trajectories of potential outcomes** $Y_t(\overline{\zeta}_t{}^{+})$ and $Y_t(\overline{\zeta}_t{}^{0})$ for the experimental dataset. This requires modeling the impact of a full sequence of treatments while adjusting for post-treatment, time-varying covariates.
2. **Imputing long-term outcomes from a short-term experiment.** This is difficult because the experiment tests a new ranker that does not exist in the observational dataset.
3. **Handling covariate shifts** when borrowing information from the observational data. Not accounting for seasonal patterns or changes in user demographics between the experimental and observational data can lead to biased treatment effect estimates.

Our Proposal

We address each of the estimation challenges by synthesizing three methodological techniques into one unified pipeline, which we call `BSTAR` (**B**lip **S**urrogate **Tr**A**nsfe**R**):

1. ***Blips from structural nested mean models***. To solve Challenge 1, we borrow the concept of blip functions from structural nested mean models (SNMMs) [5]. Instead of directly modeling the longitudinal potential outcomes as a complex function of history, we view them as an accumulation of "blips" or incremental impacts of treatments across time points. Concretely, the blip functions are defined for $t \geq r$ as follows:

$$\gamma_{r,t}(\overline{Z}_r, \overline{S}_{r-1}, \overline{X}_{r-1}; \zeta, d) \triangleq \mathbb{E}\left\{Y_t(\overline{Z}_r, \underline{\zeta}_{r+1}) - Y_t(\overline{Z}_{r-1}, \underline{\zeta}_r) \mid \overline{Z}_r, \overline{S}_{r-1}, \overline{X}_{r-1}, D = d\right\}.$$

   $\gamma_{r,t}(\cdot)$ measures the point-in-time effect that the treatment at time r has on the outcome at time t, after adjusting for past treatments and covariates. Using these blip functions, we can recursively identify the conditional mean of the potential outcomes. Let $U_{r,t}(\zeta, d) = Y_t - \sum_{s=r}^{t} \gamma_{s,t}(\overline{Z}_t, \overline{S}_{t-1}, \overline{X}_{t-1}; \zeta, d)$ be the potential outcome at time t with treatment effects from times r through t removed. The following equation holds:

$$\mathbb{E}\left\{U_{r,t}(\zeta, d) \mid \overline{Z}_r = \overline{z}_r, \overline{S}_r, \overline{X}_{r-1}, D = d\right\} = \mathbb{E}\left\{Y_t(\overline{z}_{r-1}, \underline{\zeta}_r) \mid \overline{Z}_r = \overline{z}_r, \overline{S}_r, \overline{X}_{r-1}, D = d\right\}.$$

   This use of SNMMs allows us to identify effects of sustained ranker changes while adjusting for time-varying confounders and dependencies between past and future ranker exposure.

2. ***Surrogate index modeling.*** To solve Challenge 2, we use the idea of surrogate index modeling [1]. We assume the *search result set* (i.e., collection of results shown by a ranker) is a surrogate variable that mediates ranker effects on target outcomes (e.g., making a purchase). This is plausible because users are blind to backend ranker algorithm details; the way any user experiences a ranker is through what is displayed on the search results page. It is this surrogacy assumption that enables information synthesis across rankers, as the result sets give a common measure to align outputs from different ranking algorithms and architectures. Mathematically, we assume:

$$\mathbb{E}\left\{Y_t(\overline{Z}_r, \underline{\zeta}_{r+1}) \mid \overline{Z}_r, \overline{S}_r, \overline{X}_{r-1}, D = d\right\} = \mathbb{E}\left\{Y_t(\overline{Z}_r, \underline{\zeta}_{r+1}) \mid \overline{S}_r, \overline{X}_{r-1}, D = d\right\}. \tag{SI}$$

   This states that the potential outcomes are mediated by the result sets output by a ranker, and are conditionally independent of the rankers. Assumption (SI) allows the following identification formula for the blip functions: For a given ranker ζ and a given population $d \in \{o, e\}$, we have

$$\gamma_{r,t}(\overline{Z}_r, \overline{S}_{r-1}, \overline{X}_{r-1}; \zeta, d) = \mathbb{E}\left\{g_{r+1,t}(\overline{X}_{r-1}, \overline{S}_r; \zeta, d)\{1 - w(\overline{Z}_r, \overline{S}_r, \overline{X}_{r-1}; \zeta)\} \mid \overline{Z}_r, \overline{S}_{r-1}, \overline{X}_{r-1}, D = d\right\},$$

   where $g_{r+1,t}(\cdot)$ is a function that predicts blipped potential outcome $U_{r+1,t}(\zeta)$ based on result set and user covariates, and $w(\cdot)$ is a density ratio that compares how two rankers generate result sets based on user information:

$$g_{r+1,t}(\overline{S}_r, \overline{X}_{r-1}; \zeta, d) = \mathbb{E}\left\{U_{r+1,t}(\zeta) \mid \overline{S}_r, \overline{X}_{r-1}, D = d\right\}, \quad w(\overline{Z}_r, \overline{S}_r, \overline{X}_{r-1}; \zeta) = \frac{p(S_r \mid \overline{Z}_{r-1}, \overline{S}_{r-1}, \overline{X}_{r-1}, Z_r = \zeta)}{p(S_r \mid \overline{Z}_{r-1}, \overline{S}_{r-1}, \overline{X}_{r-1}, Z_r)}.$$

3. ***Transfer learning.*** To solve Challenge 3, we adopt generalizability and transportability approaches from the causal transfer learning literature [3].Concretely, we make the transferability assumption:

$$\mathbb{E}\left\{Y_t(\overline{\zeta}_T) \mid X_0, D = o\right\} = \mathbb{E}\left\{Y_t(\overline{\zeta}_T) \mid X_0, D = e\right\}. \tag{TR}$$

   This states that the conditional mean of potential outcomes under a given ranker are transferable across the observational and experimental populations. Assuming transferability conditional on covariates allows us to handle covariate shifts, while heterogeneity across time points lets us capture seasonal patterns. In practice, this means we first build a causal model on the observational data using a relevant time period (e.g., season) of data, then transfer the learned causal model to the experiment by marginalizing over the experimental data's covariates.

By integrating blips, surrogate index modeling, and transfer learning into a unified pipeline, `BSTAR` allows us to identify and estimate PLATE (1).

Theorem 1 (Identification of PLATE)**.** *Under Assumptions* (SR), (SI), (TR)*, the prospective long-term average treatment effect can be identified as follows:* $\tau = \mu(\zeta^+) - \mu(\zeta^0)$ *where*

$$\mu(\zeta) = \mathbb{E}\left\{\mathbb{E}\left\{\sum_{t=1}^{T}\left(Y_t - \sum_{r=1}^{t} g_{r+1,t}(\overline{S}_r, \overline{X}_{r-1}; \zeta, o)\{1 - w(\overline{Z}_r, \overline{S}_r, \overline{X}_{r-1}; \zeta)\}\right) \mid X_0, D = o\right\}, D = e\right\} \quad \textit{for} \quad \zeta = \zeta^+, \zeta^0$$

Theorem 1 translates into an estimation pipeline, outlined in Figure 1 and described in more detail in Appendix A.

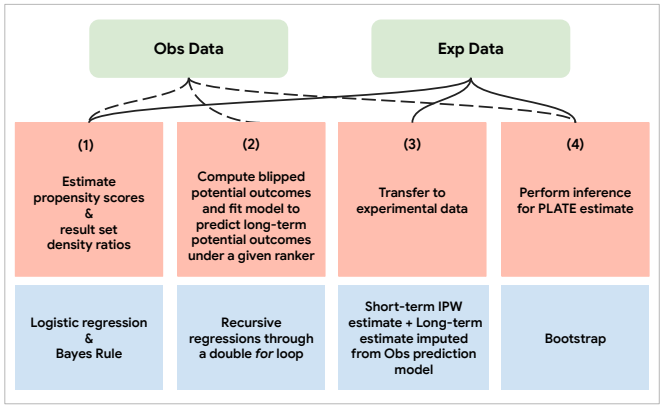

Figure 1: Overview of the estimation and inference pipeline. Red boxes (middle row) give a high-level description of the 4 main steps in the pipeline. Blue boxes (bottom row) call out modeling techniques used in each step. Dashed and solid connecting lines indicate which data source (Obs or Exp) is used in each step of the pipeline.

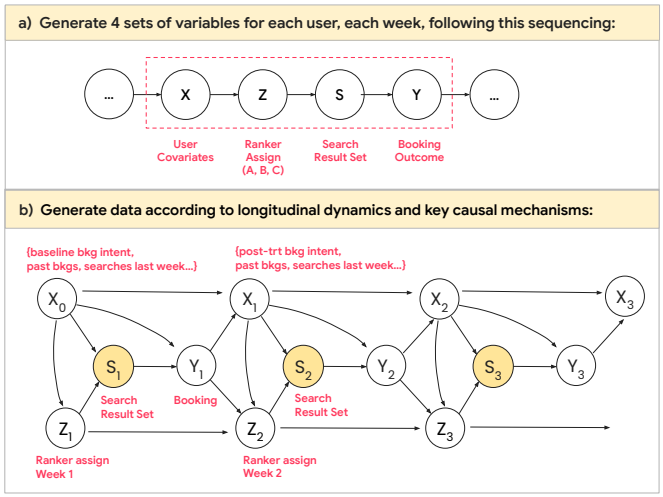

Figure 2: Visual of the simulation study dgp, designed after real ranker experiments at Airbnb. In panel (b), the colored nodes for S_t call out a key assumption enforced in the dgp: a ranker Z affects the booking outcome Y only through the search result set S.

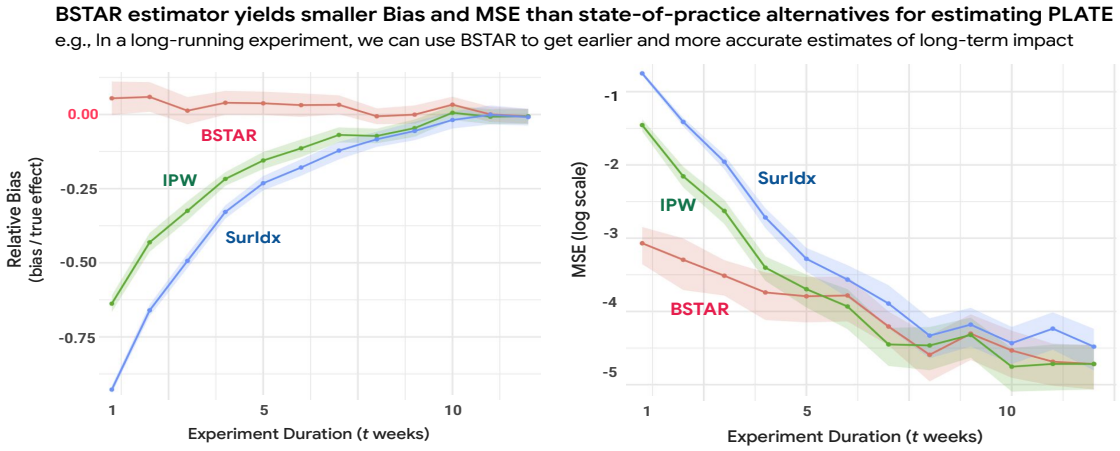

Figure 3: Simulations show `BSTAR` outperforms state-of-practice approaches `SurIdx` and `IPW` when there are complex longitudinal dynamics. We plot the bias and MSE of each approach in estimating a 12-week-long PLATE, assuming the experiment run to test this treatment (e.g., Ranker C vs Ranker A) lasts for only t weeks, with $t \in \{1, \ldots, 12\}$.

Simulation Results

We ran a set of simulations to assess the performance of `BSTAR`, using a study design motivated by real ranker experiments commonly seen in large-scale marketplaces such as Airbnb. Importantly, we generate data according to longitudinal dynamics involving sequential, time-varying treatments (e.g., ranker models) and time-varying confounders (e.g., user booking intent), and enforce a key assumption that a ranker affects the outcome of interest (e.g., booking) only through the search result set. Fig. 2 illustrates this data generating process. Under these dynamics, Fig. 3 shows that `BSTAR` yields smaller bias and MSE than simpler state-of-practice estimators, including a simple inverse propensity weighted difference-in-outcome-means estimator (`IPW`) and surrogate index estimator based on [1] (`SurIdx`).[1] Compared to these state-of-practice approaches, `BSTAR` gives more accurate estimates of the 12-week PLATE using shorter-run experiments, by properly accounting for longitudinal dynamics involving sequential, time-varying treatments and time-varying confounders. Unsurprisingly, sensitivity analyses showed that `BSTAR` does give biased estimates if important time-varying confounders (e.g., user booking intent) are omitted. For practical applications, we must rely on expert-informed and AI-assisted feature engineering to create rich enough data summaries of variables needed to adjust for confounding.

[1] The `SurIdx` approach fits a surrogate index model on the historical observational data to predict long-term cumulative outcome as a function of short-term cumulative outcome and user covariates, without further adjusting for ranker exposure or search result sets. It then calculates the difference in mean predicted cumulative long-term outcome for the experiment's treatment vs control users.

References

[1] Susan Athey, Raj Chetty, Guido Imbens, and Hyunseung Kang. 2019. Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index. *arXiv preprint arXiv:1603.09326* (2019).

[2] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Miruna Oprescu, and Vasilis Syrgkanis. 2022. Estimating the Long-Term Effects of Novel Treatments. arXiv:2103.08390 [econ.EM] `https://arxiv.org/abs/2103.08390`

[3] Irina Degtiar and Sherri Rose. 2023. A Review of Generalizability and Transportability. *Annual Review of Statistics and Its Application* 10, Volume 10, 2023 (2023), 501–524. doi:`10.1146/annurev-statistics-042522-103837`

[4] Tianchen Qian, Hyesun Yoo, Predrag Klasnja, Daniel Almirall, and Susan A Murphy. 2021. Estimating time-varying causal excursion effects in mobile health with binary outcomes. *Biometrika* 108, 3 (2021), 507–527.

[5] James M Robins. 2004. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics: analysis of correlated data*. Springer, 189–326.

[6] Allen Tran, Aurélien Bibaut, and Nathan Kallus. 2024. Inferring the Long-Term Causal Effects of Long-Term Treatments from Short-Term Experiments. arXiv:2311.08527 [stat.AP] `https://arxiv.org/abs/2311.08527`

A Detailed algorithm for PLATE estimation and inference

Theorem 1 translates into an estimation algorithm:

1. On experimental and observational data:
   (a) Estimate the propensity scores $p(Z_r \mid \overline{Z}_{r-1}, \overline{S}_{r-1}, \overline{X}_{r-1}, D = e)$ and $p(Z_r \mid \overline{Z}_{r-1}, \overline{S}_{r-1}, \overline{X}_{r-1}, D = o)$ for the experimental and observational data, respectively.
   (b) Estimate the density ratio $w(\overline{Z}_r, \overline{S}_r, \overline{X}_{r-1}; \zeta)$ for $\zeta = \zeta^0, \zeta^+$ using both the experimental and observational data.
2. On observational data:
   (a) Define $U_{t+1,t,i}(\zeta) = Y_{t,i}$, for $t = 1, \ldots, T_{\text{LONG}}$ and $\zeta = \zeta^0, \zeta^+$ for $i = 1, \ldots, N_o$.
   (b) Perform recursive blipping through a double-layer loop. For each $r = T, \ldots, 1$, $t = T, \ldots, r$ and $\zeta = \zeta^0, \zeta^+$, do the loop:
      i. Estimate the function $g_{r+1,t}(\overline{S}_r, \overline{X}_{r-1}; \zeta, o)$ by regressing $U_{t+1,t,i}(\zeta)$ on $\overline{S}_{r,i}$ and $\overline{X}_{r-1,i}$.
      ii. Estimate the shift function $\delta_{r+1,t}(\overline{Z}_r, \overline{S}_{r-1}, \overline{X}_{r-1}; \zeta)$ for demeaning using (1.1):
      $$\delta_{r+1,t}(\overline{Z}_r, \overline{S}_{r-1}, \overline{X}_{r-1}; \zeta) = \mathbb{E}\left\{g_{r+1,t}(\overline{X}_{r-1}, \overline{S}_r; \zeta, d) \mid \overline{Z}_r, \overline{S}_{r-1}, \overline{X}_{r-1}, D = d\right\}. \tag{1.1}$$
      iii. Construct reweighted difference $\Delta_{r,t,i}(\zeta)$ using (1.2):
      $$\Delta_{r,t}(\zeta) = \{g_{r+1,t}(\overline{S}_r, \overline{X}_{r-1}; \zeta, o) - \delta_{r+1,t}(\overline{Z}_r, \overline{S}_{r-1}, \overline{X}_{r-1}; \zeta)\}\{1 - w(\overline{Z}_r, \overline{S}_r, \overline{X}_{r-1}; \zeta)\}, \tag{1.2}$$
      iv. Estimate the blip function $\gamma_{r,t}(\overline{Z}_r, \overline{S}_{r-1}, \overline{X}_{r-1}; \zeta, d)$ by regressing $\Delta_{r,t,i}(\zeta)$ on $\overline{Z}_{r,i}$, $\overline{S}_{r-1,i}$, and $\overline{X}_{r-1,i}$.
      v. Update $U_{r,t,i}(\zeta) = U_{r+1,t,i}(\zeta) - \gamma_{r,t}(\overline{Z}_{r,i}, \overline{S}_{r-1,i}, \overline{X}_{r-1,i}; \zeta, d)$.
   (c) Impute total cumulative potential outcome for the i-th unit on observational data if they were always using ranker $\zeta = \zeta^0, \zeta^+$:
   $$\widetilde{Y}_{(T_{\text{SHORT}}+1):T_{\text{LONG}},i}(\zeta, o) = \sum_{t=T_{\text{SHORT}}+1}^{T_{\text{LONG}}} U_{1,t,i}(\zeta).$$
   (d) Fit prediction model $\mu_{(T_{\text{SHORT}}+1):T_{\text{LONG}}}(X_0; \zeta)$ by regressing $\widetilde{Y}_{(T_{\text{SHORT}}+1):T_{\text{LONG}},i}(\zeta, o)$ on $X_{0,i}$ on the observational data for $\zeta = \zeta^0, \zeta^+$.
3. On experimental data:
   (a) Use IPW or AIPW to impute the short-term cumulative potential outcomes: $\widetilde{Y}_{1:T_{\text{SHORT}},i}(\zeta, e)$ for $i = 1, \ldots, N_e$.
   (b) Impute the missing long-term cumulative potential outcomes: for $\zeta = \zeta^0, \zeta^+$,
   $$\widetilde{Y}_{(T_{\text{SHORT}}+1):T_{\text{LONG}},i}(\zeta, e) = \sum_{t=T_{\text{SHORT}}+1}^{T_{\text{LONG}}} \mu_{(T_{\text{SHORT}}+1):T_{\text{LONG}}}(X_{0,i}; \zeta).$$
   (c) Impute the total cumulative potential outcomes:
   $$\widetilde{Y}_{1:T_{\text{LONG}},i}(\zeta, e) = \widetilde{Y}_{1:T_{\text{SHORT}},i}(\zeta, e) + \widetilde{Y}_{(T_{\text{SHORT}}+1):T_{\text{LONG}},i}(\zeta, e).$$

(d) Compute the long-term effects: $\widehat{\tau} = N_e^{-1} \sum_{i=1}^{N_e} \widetilde{Y}_{1:T_{\text{LONG}},i}(\zeta^+, e) - \widetilde{Y}_{1:T_{\text{LONG}},i}(\zeta^0, e)$.

4. On experimental and observational data: Perform inference for the PLATE estimate by bootstrapping the above steps.

B Additional Simulation Results

To better understand the statistical properties of `BSTAR`, we conducted additional simulations to evaluate its bias, MSE, coverage rate, and power under different sample sizes. Under a data generating process designed to match the realistic setting of ranker model experiments (as illustrated in Fig. 2), we vary the data sample size on an exponential scale and perform evaluation. The left panel of Figure 4 shows that the relative bias of `BSTAR` is small, and approaches zero as sample size increases. The reason for large bias for small sample sizes is that the result set density ratio estimation suffers from numerical stability issues when sample size is limited. The right panel of Figure 4 shows that the MSE exhibits a log-log linear trend, which verifies a polynomial convergence rate of the estimator against sample size.

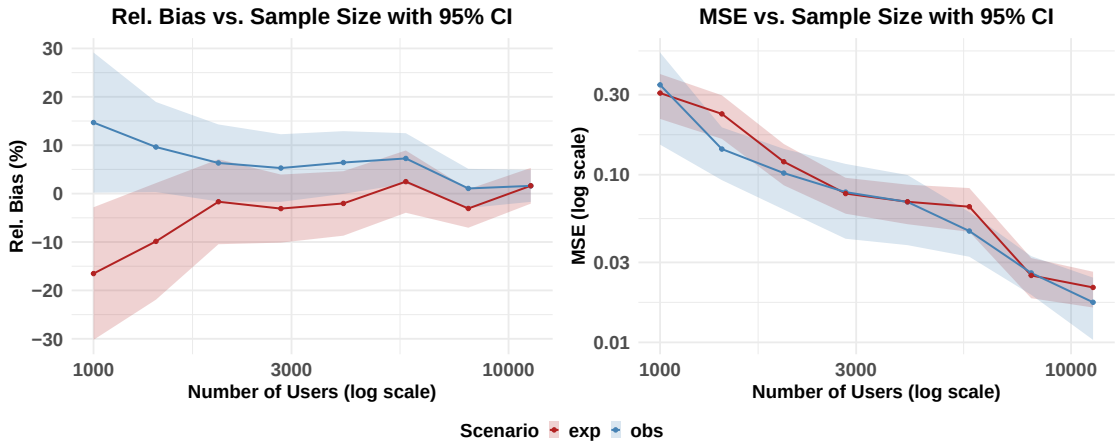

Figure 4: Bias and MSE for the `BSTAR` estimator as a function of sample size.

Figure 5 shows the coverage rate and power rate of `BSTAR` as a function of sample size. The confidence intervals are constructed using quantile-based bootstrap intervals. The coverage rates are well controlled at the 95% level and the power goes to one asymptotically as sample size goes to infinity.

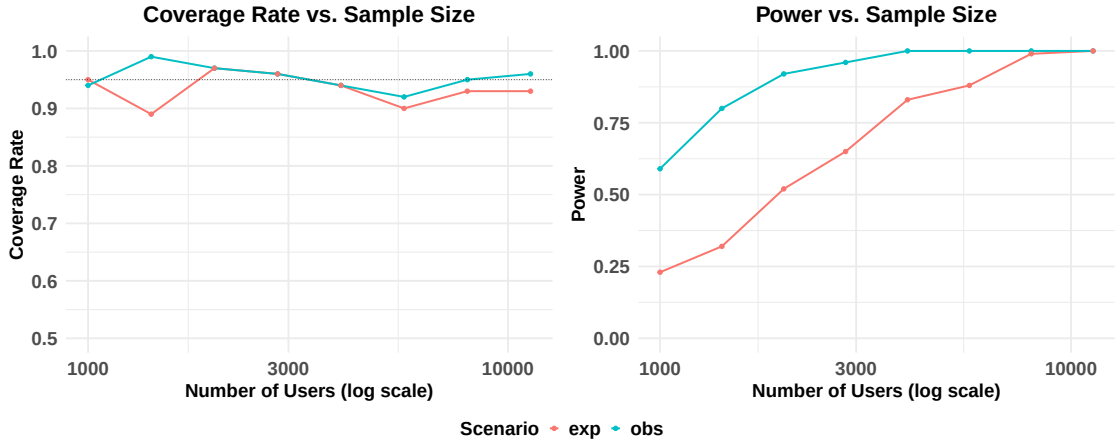

Figure 5: Coverage rates and power for `BSTAR` estimator with Bootstrap methods for variance estimation.