Post-Mortem of RL

The goal of RL was to personalize, namely, maximize cumulative rewards.

But did it achieve it?

If we use the best RL algorithm, do we need to bother?

- Theoretically, when suitable assumptions hold, a low-regret RL algorithm achieves low regret as number of decision times becomes large

- **Point 1:** While theoretical guarantees can help design an RL algorithm, there are many challenges when using them in real-life. Many changes may have to be made for practical reasons

If we use the best RL algorithm, do we need to bother?

- E.g., the HeartSteps RL
    - used simple linear model with few features (bias-variance tradeoff to speed up learning)
    - used Gaussian prior and likelihood (to obtain an autonomous, computationally stable, online algorithm)
    - estimated delayed effects on prior data and did not incorporate the uncertainty in estimating the delayed effects (bias-variance tradeoff)
    - used clipping, to manage burden and enable after-study analyses

Point 2: Even if we use the ``best RL algorithm", we *do* need to bother!

- Did the **deployed RL algorithm** actually personalize for the problem at hand
    - Namely, did it maximize rewards **within** the duration of the study **for** the users **and** the underlying (**unknown**) data generating process?

Point 2: Even if we use the ``best RL algorithm", we *do* need to bother!

- Did the **deployed RL algorithm** actually personalize for the problem at hand
    - Namely, did it maximize rewards **within** the duration of the study **for** the users **and** the underlying (**unknown**) data generating process?
- It may **not** have done well because of low number of decision times, low amount of availability, misspecified modeling and high noise in the data.

Can we use a proxy to see if RL is maximizing rewards?

- As a starting point, we can look directly at the estimates of the **advantage** forecasts by the RL algorithm

- **Advantage** of action 1 over action 0 in state s
  $= r(s,1) - r(s,0) + \gamma\left(H^{\pi^\star}(s,1) - H^{\pi^\star}(s,0)\right) =$ Treatment effect + Delayed effect

Can we use a proxy to see if RL is maximizing rewards?

- As a starting point, we can look directly at the estimates of the **advantage** forecasts by the RL algorithm

- **Advantage** of action 1 over action 0 in state s
  $= r(s,1) - r(s,0) + \gamma\left(H^{\pi^\star}(s,1) - H^{\pi^\star}(s,0)\right) =$ Treatment effect + Delayed effect

- In HeartSteps RL, we estimate

  - treatment effect $r(s,1) - r(s,0)$ in an online manner using a linear model and

  - negative delayed effect $\eta(s) \triangleq -\gamma(H^{\pi^\star}(s,1) - H^{\pi^\star}(s,0))$ using prior data

Can we use a proxy to see if RL is maximizing rewards?

- As a starting point, we can look directly at the estimates of the **advantage** forecasts by the RL algorithm

- **Advantage** of action 1 over action 0 in state s
  $= r(s,1) - r(s,0) + \gamma\left(H^{\pi^\star}(s,1) - H^{\pi^\star}(s,0)\right) =$ Treatment effect + Delayed effect

- In HeartSteps RL, we estimate

  - treatment effect $r(s,1) - r(s,0)$ in an online manner using a linear model and

  - negative delayed effect $\eta(s) \triangleq -\gamma(H^{\pi^\star}(s,1) - H^{\pi^\star}(s,0))$ using prior data

Why care about advantages?

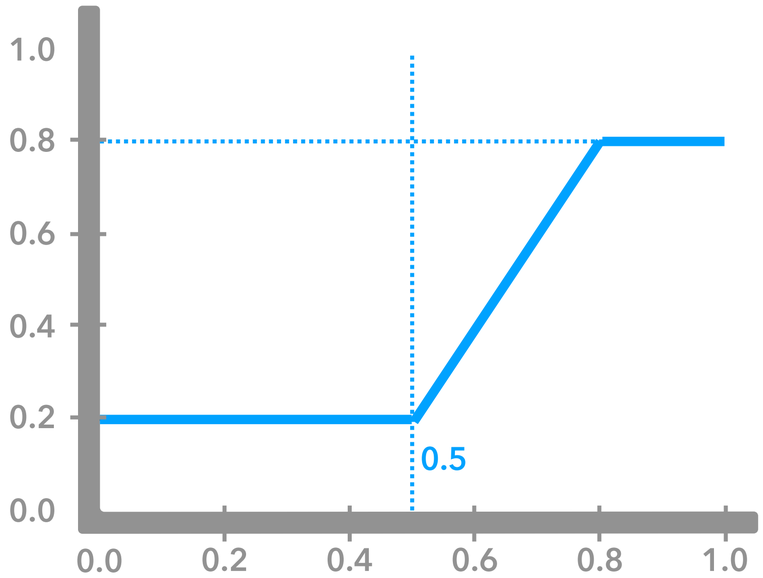
HeartSteps RL: Bayesian Thompson Sampling with Gaussian linear model

- Model for treatment effect: $r(s,1) - r(s,0) = \beta^\top f(s)$
- Let $\hat{\beta}_t, \hat{\Sigma}_t$ denote the posterior mean and variance for the treatment effect parameters on day t, and $\hat{\eta}_t$ denote the estimate for the delayed effect
- Then posterior mean and variance of advantage at time t in state s are respectively $\hat{\beta}_t^\top f(s) - \hat{\eta}_t(s)$ and $f(s)^T \hat{\Sigma}_t f(s)$

Why care about standardized advantage?

- Recall the action selection was done with a clipped version of the posterior probability

Randomization probability



Posterior probability

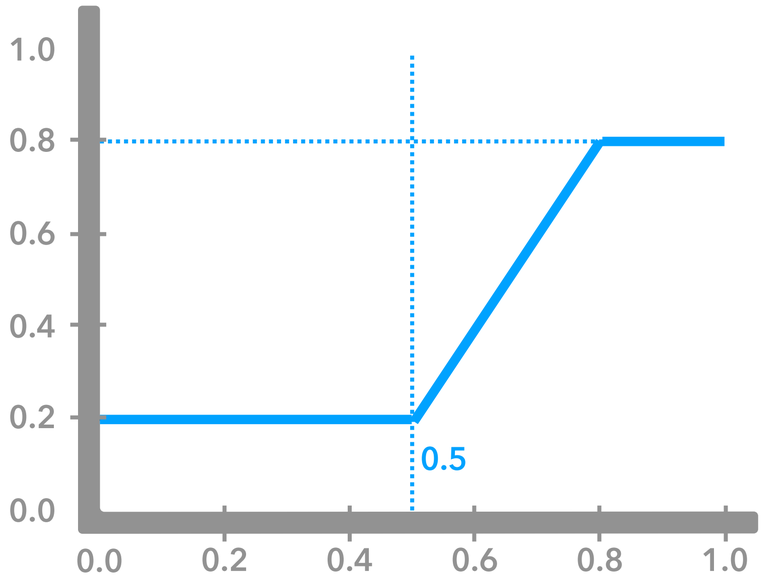
$$= \Phi(\hat{\Delta}_t(s))$$

Why care about standardized advantage?

- Recall the action selection was done with a clipped version of the posterior probability

- The posterior probability at time t in state s is $\mathbb{P}\left[\mathcal{N}(\hat{\beta}_t^\top f(s), f(s)^\top \hat{\Sigma}_t f(s)) > \hat{\eta}_t(s)\right]$

Randomization probability



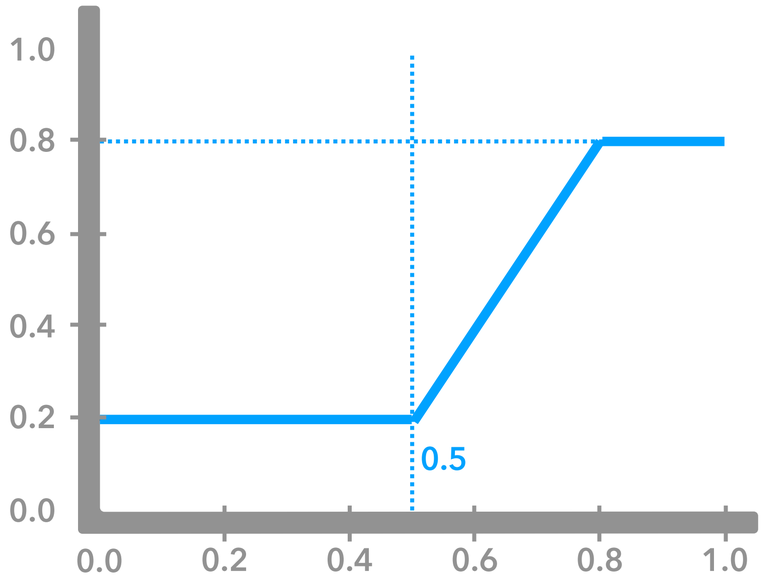
Posterior probability

$= \Phi(\hat{\Delta}_t(s))$

Why care about standardized advantage?

- Recall the action selection was done with a clipped version of the posterior probability
- The posterior probability at time t in state s is $\mathbb{P}\left[\mathcal{N}(\hat{\beta}_t^\top f(s), f(s)^\top \hat{\Sigma}_t f(s)) > \hat{\eta}_t(s)\right]$
- Which is equal to $\Phi(\hat{\Delta}_t(s))$, where $\hat{\Delta}_t(s) = \left(\hat{\beta}_t^\top f(s) - \hat{\eta}_t(s)\right) / \sqrt{f(s)^T \hat{\Sigma}_t f(s)}$ is the standardized advantage and $\Phi(x) = \mathbb{P}(\mathcal{N}(0,1) > x)$ is the inverse CDF of standard Gaussian and

Randomization probability



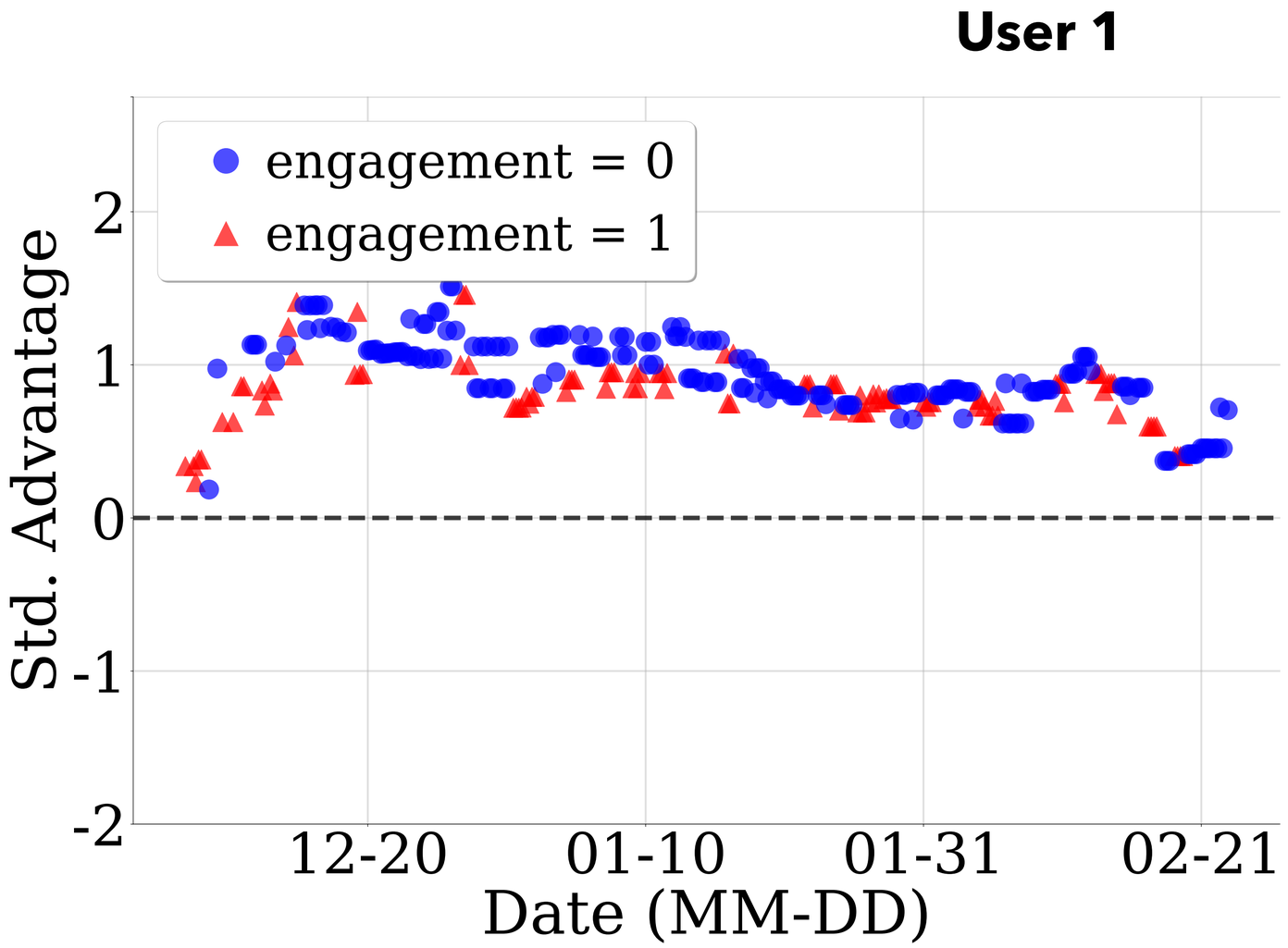
Posterior probability

$= \Phi(\hat{\Delta}_t(s))$

Features used in the linear model of HeartSteps RL

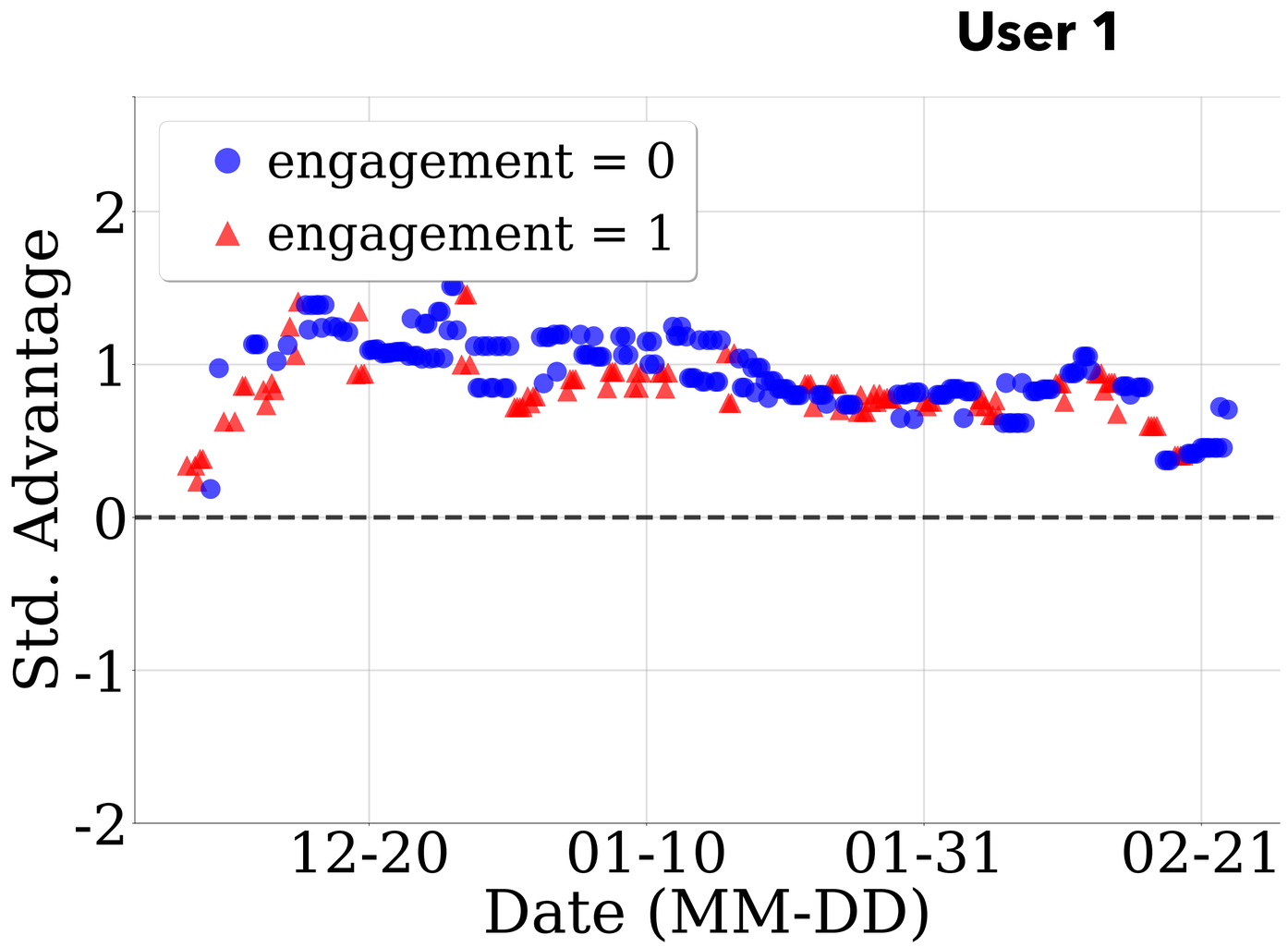
$f(s)$ has 5 components including intercept and

- Dosage [0, 20]: Discounted total number of notifications sent
- Engagement {0, 1}: Whether the user has been recently interacting with the app (1) or not (0)
- Variation {0, 1}: Whether the user's recent step counts have been highly variable (1) or not (0)
- Location {0, 1}: Whether at home/work (0) or not (1)



What is happening in this graph?

Is there an interesting trend in what actions is RL likely to recommend and when?

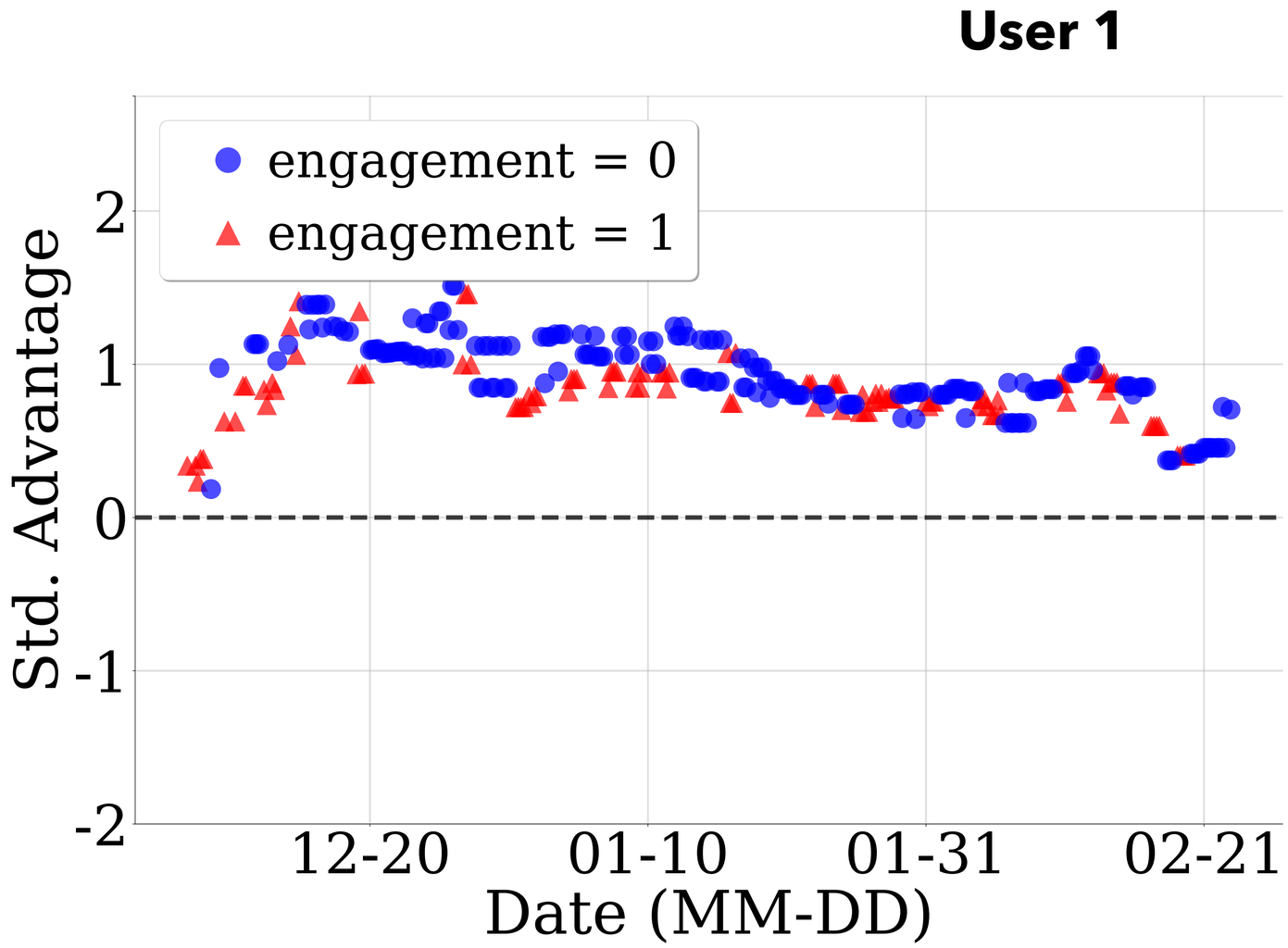


What is happening in this graph?

Is there an interesting trend in what actions is RL likely to recommend and when?

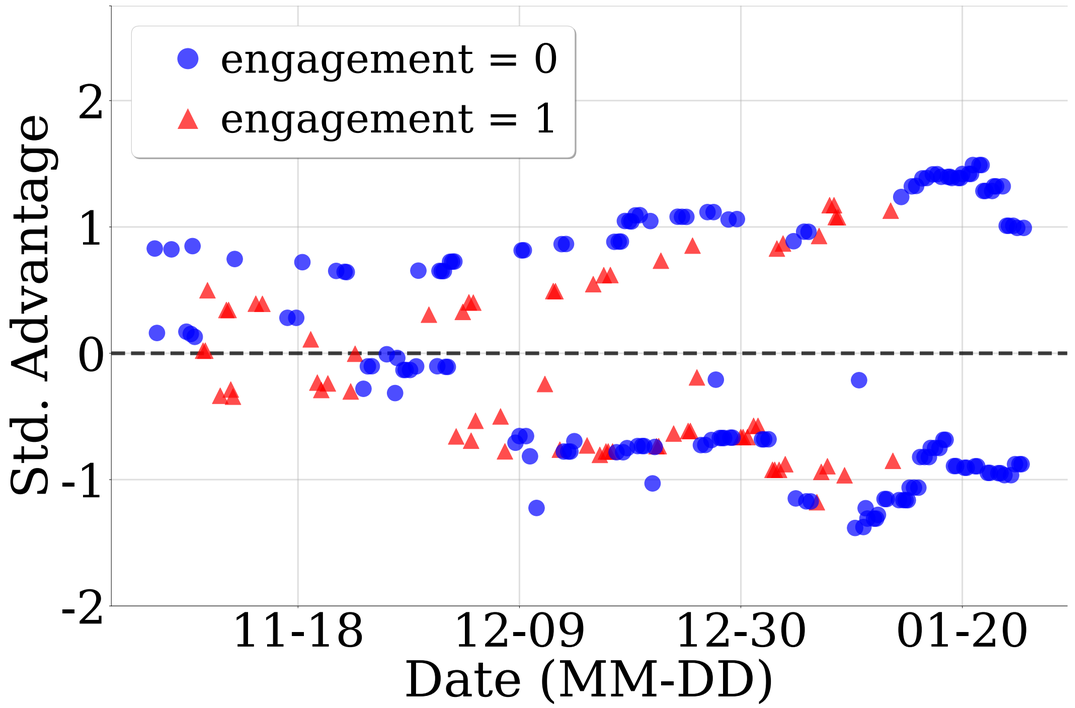
The advantage forecasts stay consistently above zero after the initial few days, regardless of state.

Did the RL algorithm personalize for this user?



by learning that sending activity suggestions is always beneficial in the states experienced by this user

User 2

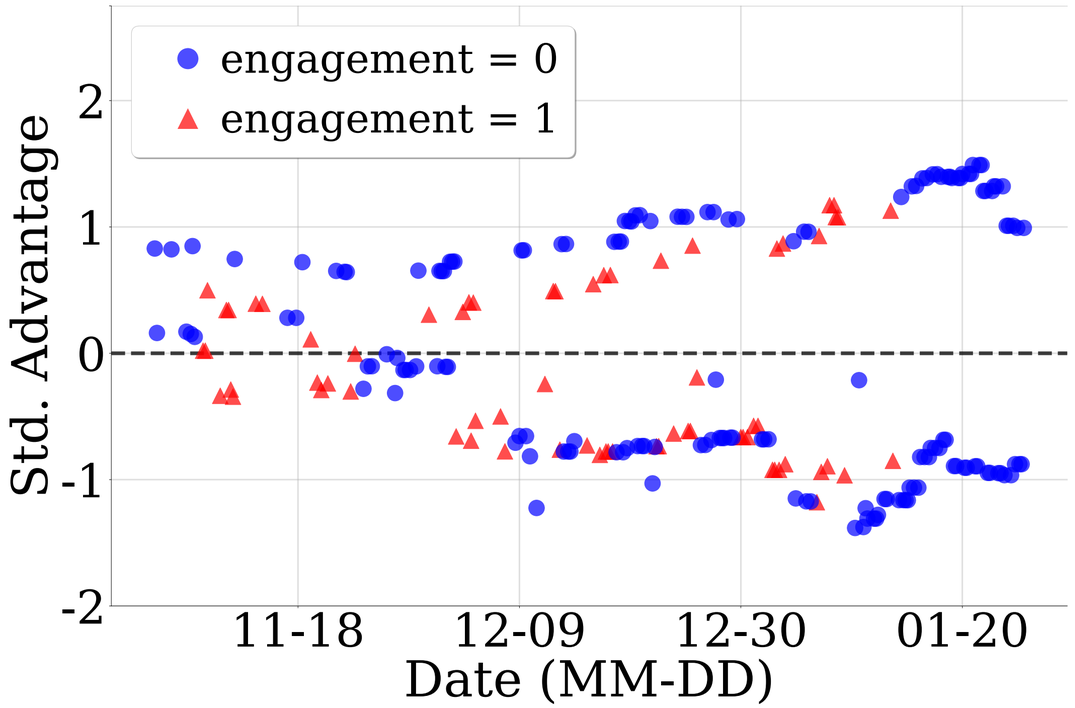


What is happening
in this graph?

Is there an interesting trend in
what actions is RL likely to
recommend and when?

Does the following graph exhibit some sort of personalization by the RL algorithm?

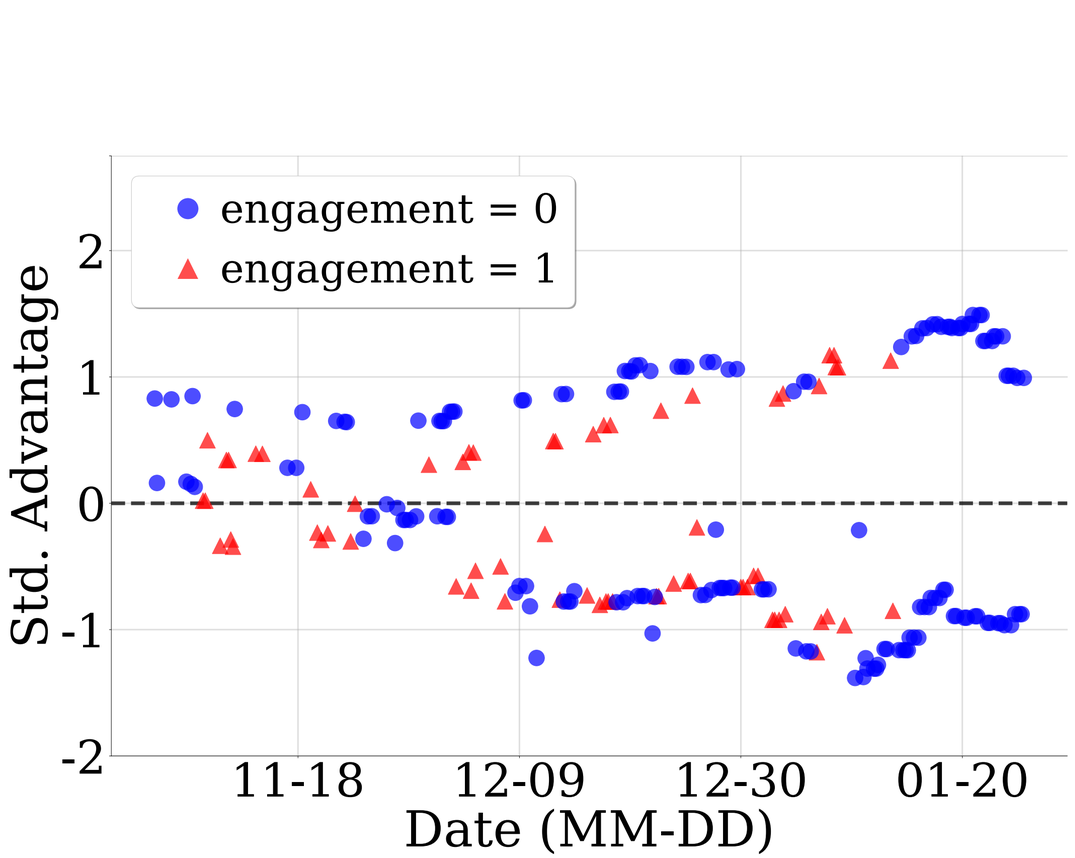
User 2



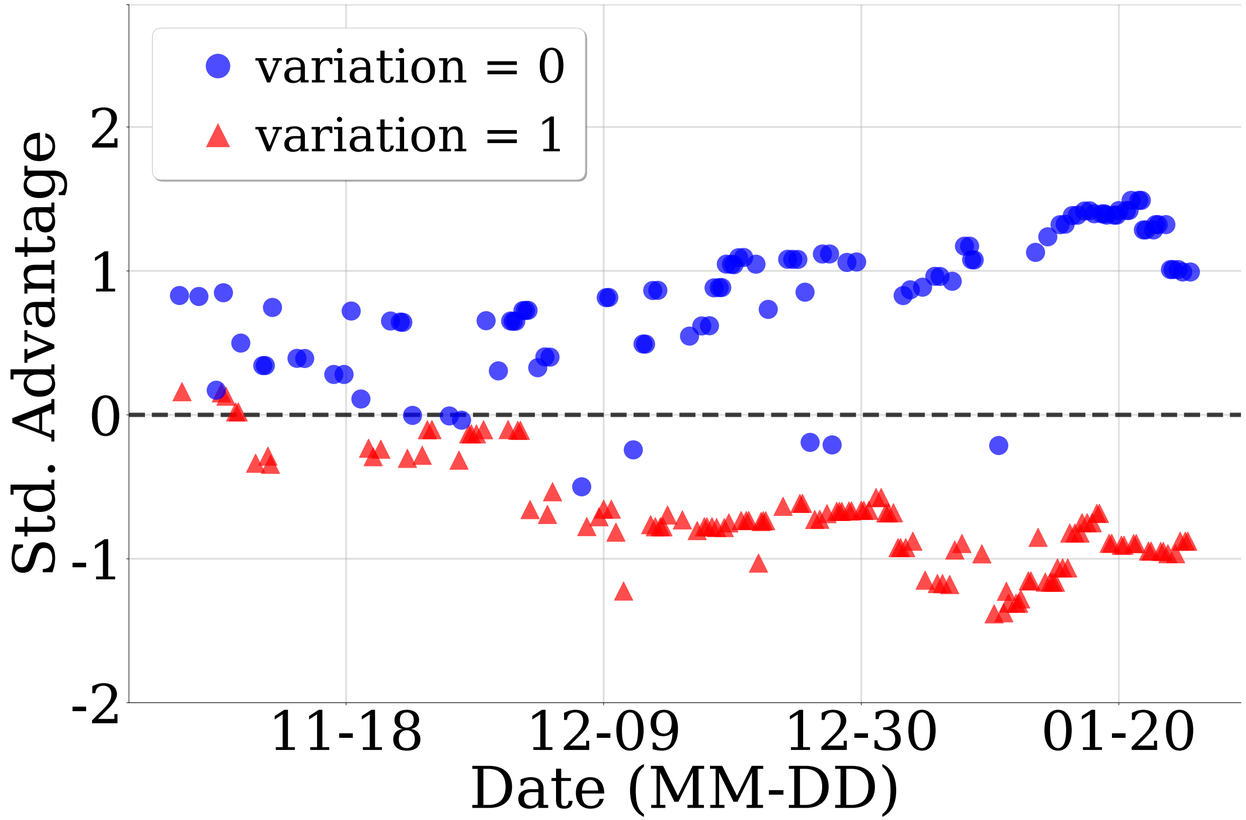
What is happening in this graph?

Is there an interesting trend in what actions is RL likely to recommend and when?

How about the figure on the right?



User 2



same forecasts, just color coded by different feature values

Demonstrating benefits of RL algorithm in real-life

- What kinds of claims could be useful to argue that RL algorithm is useful for the application at hand?

Demonstrating benefits of RL algorithm in real-life

- What kinds of claims could be useful to argue that RL algorithm is useful for the application at hand?

- Here, we investigate two kinds of claims assuming there is a way to operationalize what graph is interesting from the point of personalization:

    - If I see an interesting user graph, can we say that the RL algorithm is personalizing for this user? When would we have some confidence?

Demonstrating benefits of RL algorithm in real-life

- What kinds of claims could be useful to argue that RL algorithm is useful for the application at hand?

- Here, we investigate two kinds of claims assuming there is a way to operationalize what graph is interesting from the point of personalization:

    - If I see an interesting user graph, can we say that the RL algorithm is personalizing for this user? When would we have some confidence?

    - If I see “X” number of interesting graphs, can we say that the RL algorithm is personalizing broadly for most users if ``X'' is high? When would we have some confidence?

Can we conclude “personalization” by just looking at these figures? Will that be “truthful advertising”?

- **Hint:** The quantities in these plots are forecasts produced by the RL algorithm

Can we conclude “personalization” by just looking at these figures? Will that be “truthful advertising”?

- **Hint:** The quantities in these plots are forecasts produced by the RL algorithm

- What do they have to do with reality? What if the algorithm is hallucinating or is gaming with these forecasts? What if these estimates arise simply due to stochasticity?

    - **Sources of stochasticity**:

Can we conclude “personalization” by just looking at these figures? Will that be “truthful advertising”?

- **Hint:** The quantities in these plots are forecasts produced by the RL algorithm
- What do they have to do with reality? What if the algorithm is hallucinating or is gaming with these forecasts? What if these estimates arise simply due to stochasticity?
    - **Sources of stochasticity**: sampling of users from the population of interest, state transitions, noise in rewards, the stochastic sampling of actions in the RL algorithm
- **Classical wisdom:** If we hunt enough, we might find something interesting just by chance. ``Don’t ask the data too much, else it will lie.'' Also, known as p-hacking.

- **In this talk, we focus only on the stochasticity due to the sampling of actions by RL algorithm**

- That is we ask, if the personalization exhibited by the RL algorithm might arise solely due to the stochasticity in the sampling of actions by the RL algorithm

    - Can generalize to include other sources of stochasticity

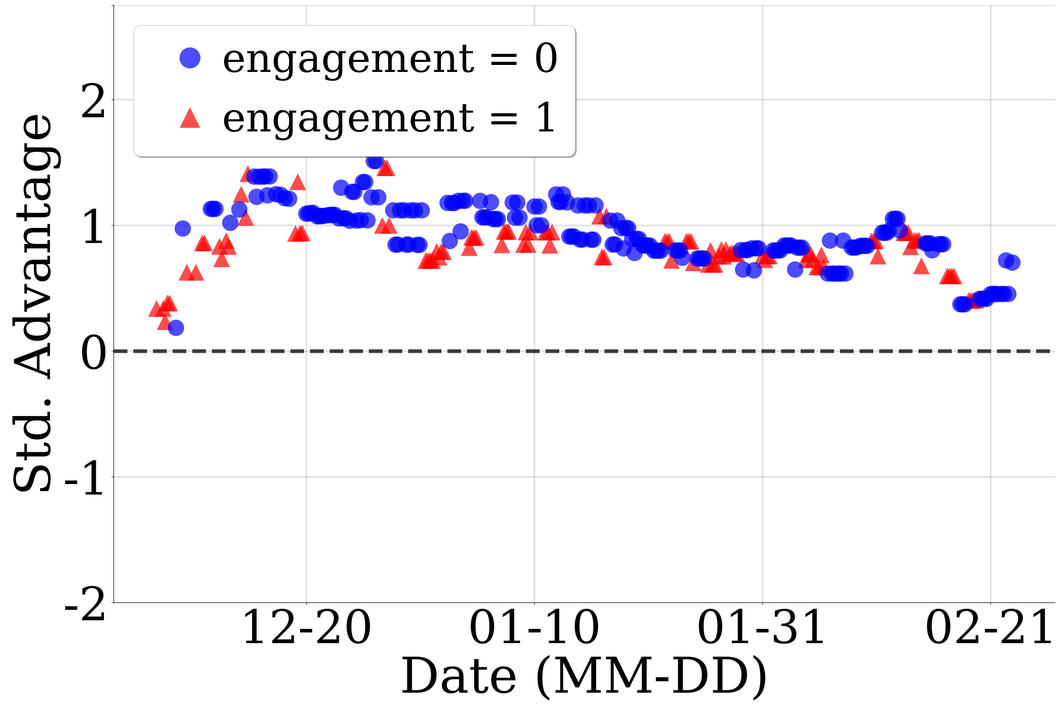
To proceed further: Operationalize personalization

- In other words, we have to quantify what are “interesting user graphs”– depends on the downstream usage / insight we want to derive
- We looked at two kinds of interestingness based on standardized advantage forecasts

Quantifying interesting graphs of type 1

Score_int$_1$ = Fraction of decision times with positive advantage forecasts

$$= \frac{|\{t : \hat{\Delta}_t(S_t) > 0\}|}{T}$$



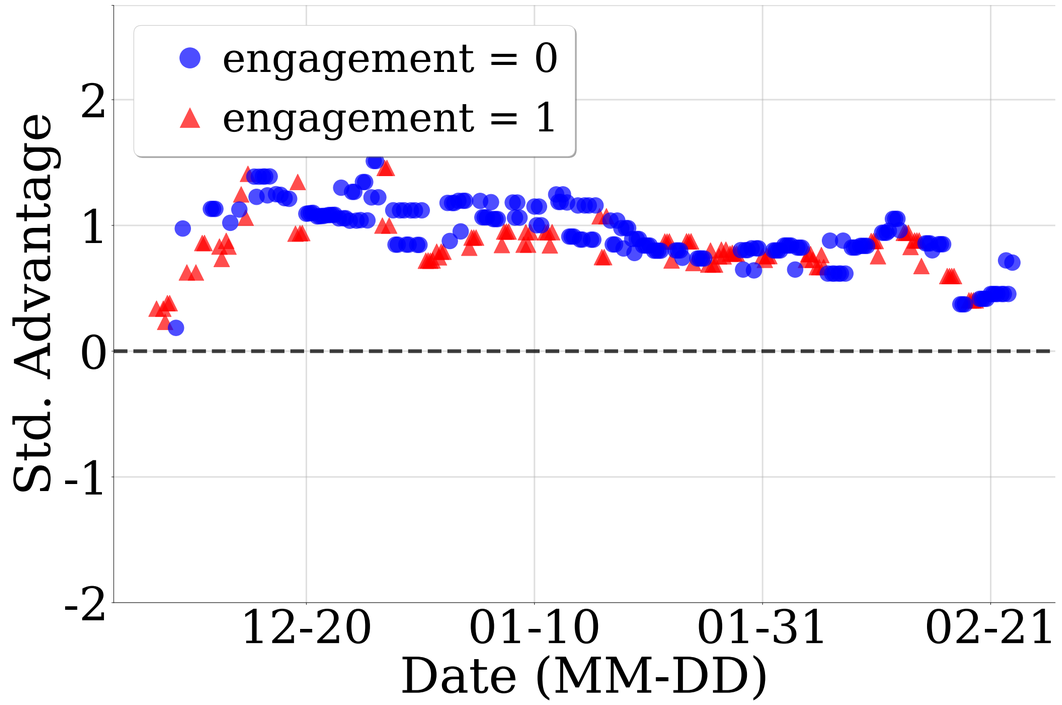
$$\hat{\Delta}_t(s) = \left(\hat{\beta}_t^\top f(s) - \hat{\eta}_t(s)\right) / \sqrt{f(s)^T \hat{\Sigma}_t f(s)}$$

Quantifying interesting graphs of type 1

Score_int$_1$ = Fraction of decision times with positive advantage forecasts

$$= \frac{|\{t : \hat{\Delta}_t(S_t) > 0\}|}{T}$$

This score takes value 1 for this user!

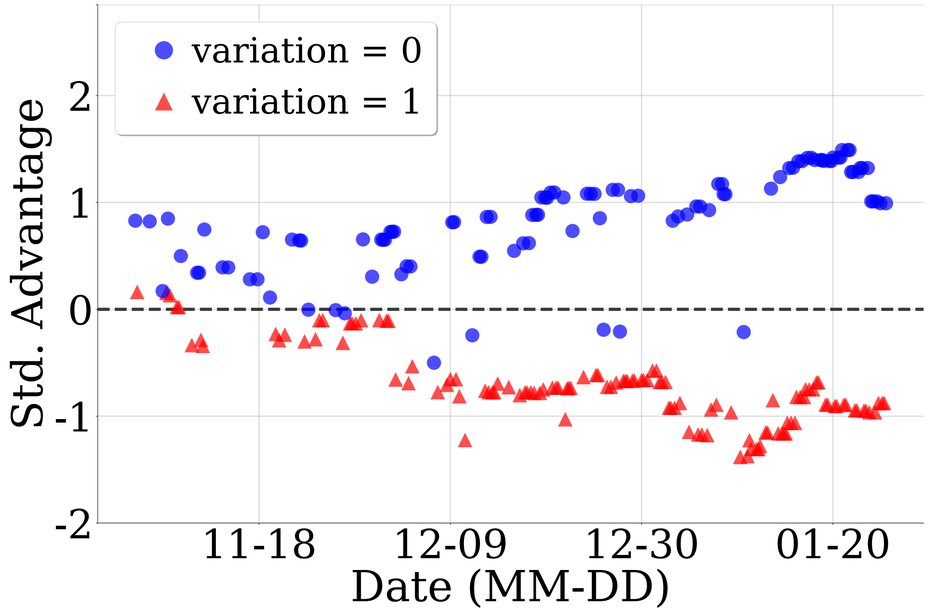


$$\hat{\Delta}_t(s) = \left(\hat{\beta}_t^\top f(s) - \hat{\eta}_t(s)\right) / \sqrt{f(s)^T \hat{\Sigma}_t f(s)}$$

Quantifying interesting graphs of type 2

Score_int$_{2,z}$ = Fraction of decision times that the advantage forecasts is higher in if a feature z takes value 1 vs 0

$$= \frac{|\{t : \hat{\Delta}_t(S_{t,z=1}) > \hat{\Delta}_t(S_{t,z=0})\}|}{T}$$



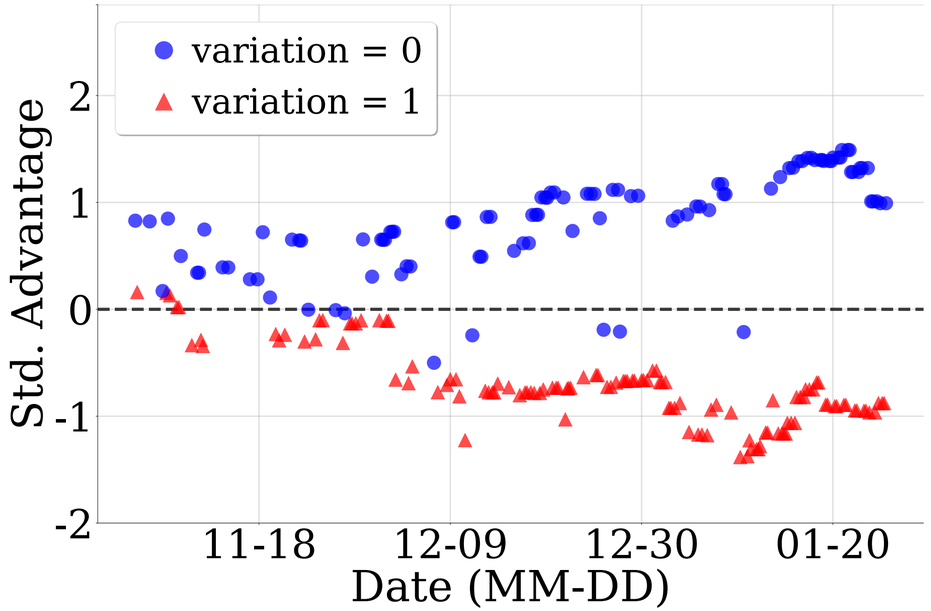
$$\hat{\Delta}_t(s) = \left(\hat{\beta}_t^\top f(s) - \hat{\eta}_t(s)\right) / \sqrt{f(s)^T \hat{\Sigma}_t f(s)}$$

Quantifying interesting graphs of type 2

Score_int$_{2,z}$ = Fraction of decision times that the advantage forecasts is higher in if a feature z takes value 1 vs 0

$$= \frac{|\{t : \hat{\Delta}_t(S_{t,z=1}) > \hat{\Delta}_t(S_{t,z=0})\}|}{T}$$

This score takes value 0 for this user for z = variation



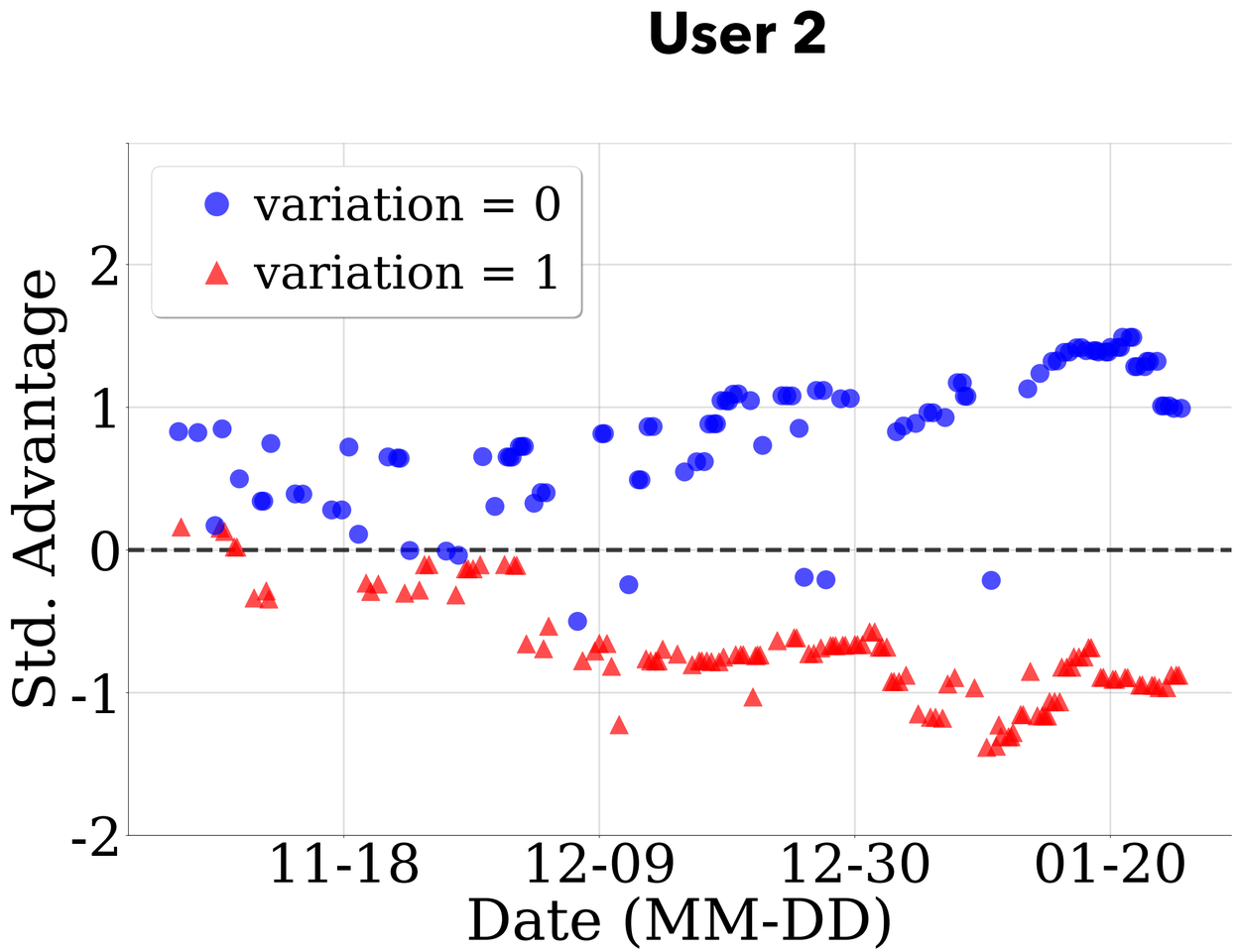
$$\hat{\Delta}_t(s) = \left(\hat{\beta}_t^\top f(s) - \hat{\eta}_t(s)\right) / \sqrt{f(s)^T \hat{\Sigma}_t f(s)}$$

Discuss

- What other ways can you think of quantifying interestingness for the two graphs? What other interestingness patterns might be useful more generally?

- Given a notion of interestingness, what other kinds of claims could be useful to argue that RL algorithm is useful for the application at hand?

Interestingness of type 2: Analysis



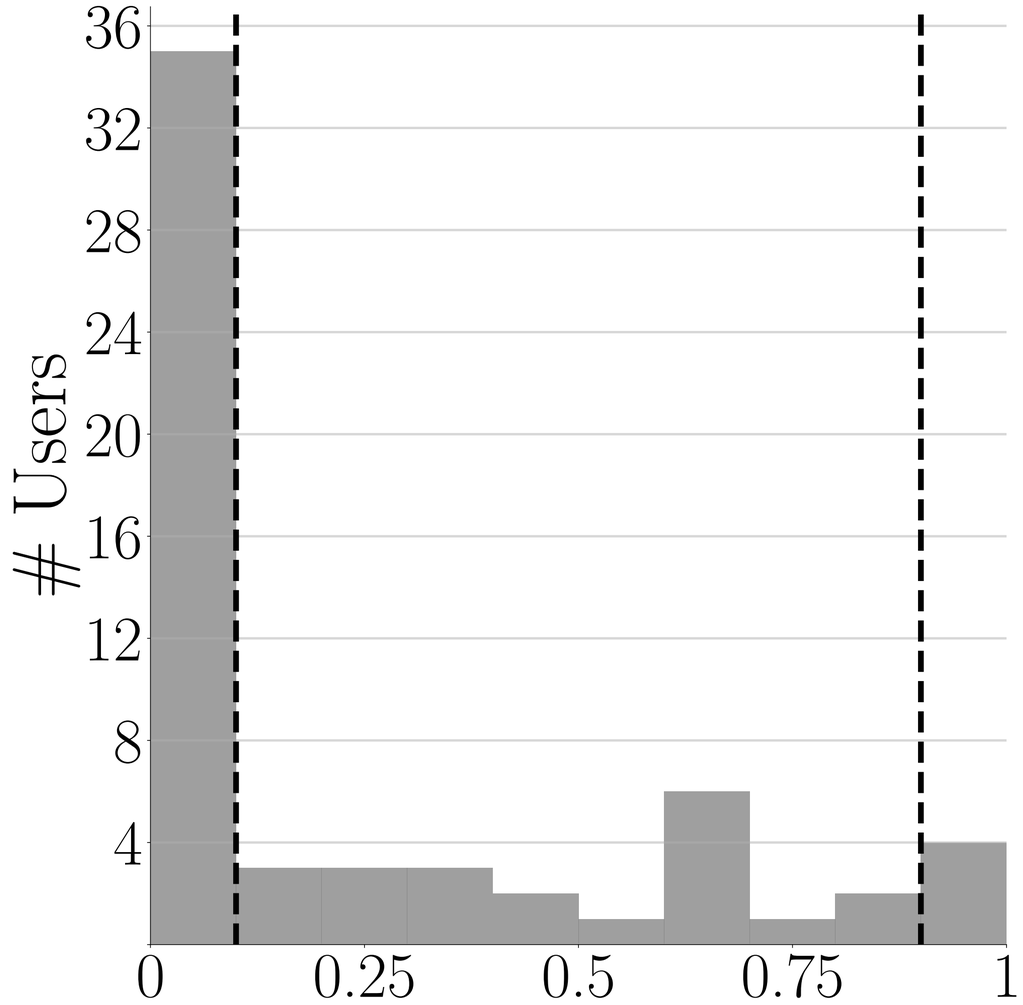
$$\text{Score\_int}_{2,z} = \frac{|\{t : \hat{\Delta}_t(S_{t,z=1}) > \hat{\Delta}_t(S_{t,z=0})\}|}{T}$$

Takes value 1 for this user

for $z = variation$

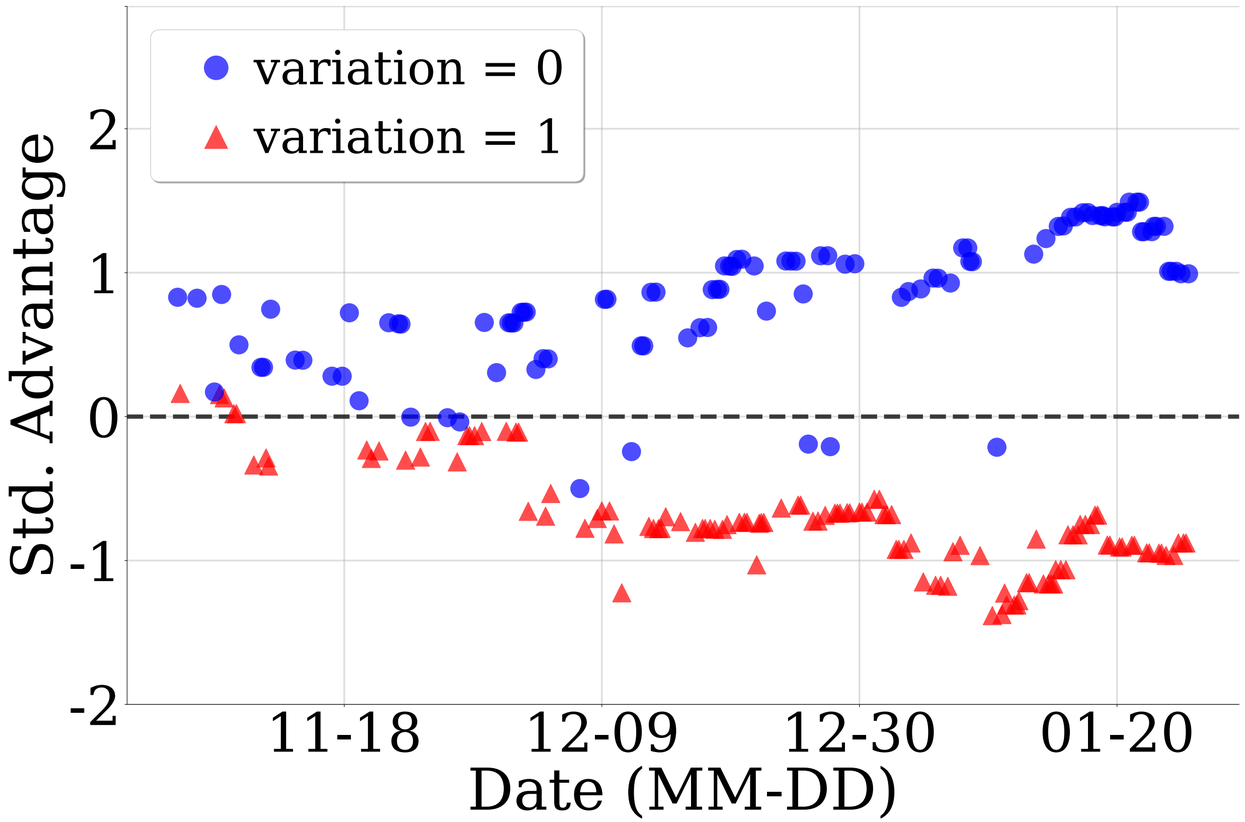
Interestingness of type 2 with $z =$ variation: Values observed in the data

- Smoothen out the forecasts across 3 days basis (at least two decision times in 3 day window each for $z = 1$ and 0)
- Filter out users with low availability
    - Why are we doing the previous two steps?
- Leaves 60 users and 39 out of these had scores $|\text{Score\_int}_{2,z} - 0.5| \geq 0.4$

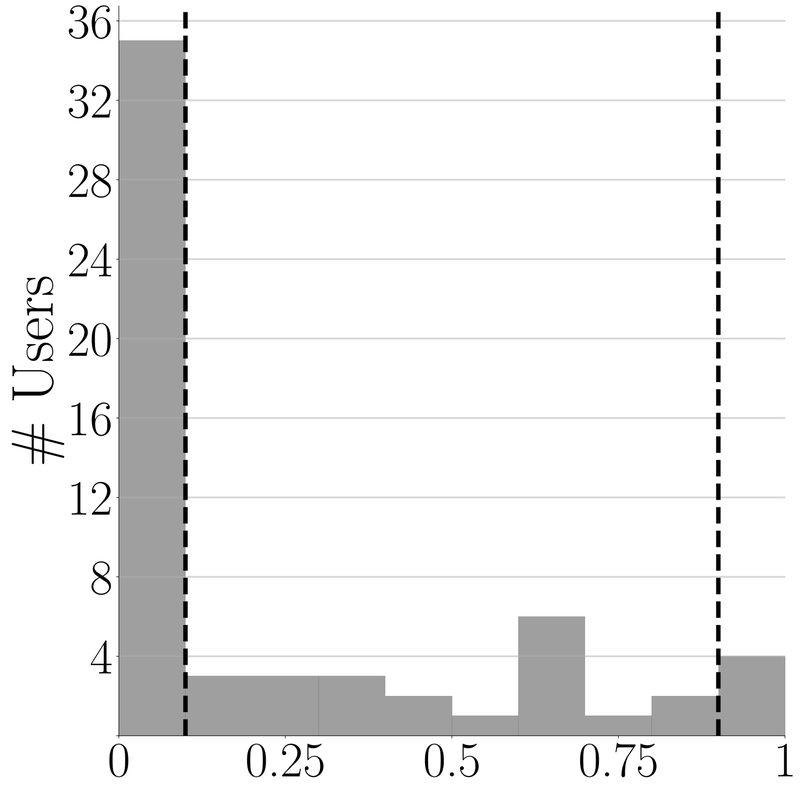


$$\text{Score\_int}_{2,z} = \frac{|\{t : \hat{\Delta}_t(S_{t,z=1}) > \hat{\Delta}_t(S_{t,z=0})\}|}{T}$$

Is RL personalizing for user 2?



Did RL learn to personalize actions based on z = "variation" for a significant number of users?



$$\text{Score\_int}_{2,z} = \frac{|\{t : \hat{\Delta}_t(S_{t,z=1}) > \hat{\Delta}_t(S_{t,z=0})\}|}{T}$$

Key idea: Check if ``personalization/interesting graphs" appear even when we know that the RL algorithm should not have personalized

Key idea: Check if ``personalization/interesting graphs" appear even when we know that the RL algorithm should not have personalized

- Have to "simulate" hypothetical scenarios where we know the ground truth and compare the trends with those in the observed data

Key idea: Check if ``personalization/interesting graphs" appear even when we know that the RL algorithm should not have personalized

- Have to "simulate" hypothetical scenarios where we know the ground truth and compare the trends with those in the observed data

- For example, would we see a graph with high Score_int$_{2,variation}$ even when there is no differential advantage of action 1 based on the value of feature variation for the users

This “simulator”-based approach requires

- A generative model to sample states, outcomes, and rewards
- Access to the RL algorithm under investigation that samples actions

Obtaining generative models

- What is the ideal generative model? As close to the “unknown” real data mechanism as possible
    - We use $r(s,a) = \alpha^\top g(s) + a\beta^\top f(s)$ for $a = 0,1$
    - Compute the posterior mean of (α, β)denoted by $\hat{\alpha}_T, \hat{\beta}_T$ for each user at final time = 90 days * 5 times/day
    - Compute residuals $\hat{\varepsilon}_t = R_t - \hat{\alpha}_T^\top g(S_t) + A_t \hat{\beta}_T^\top f(S_t)$

Resimulating/resampling user trajectory

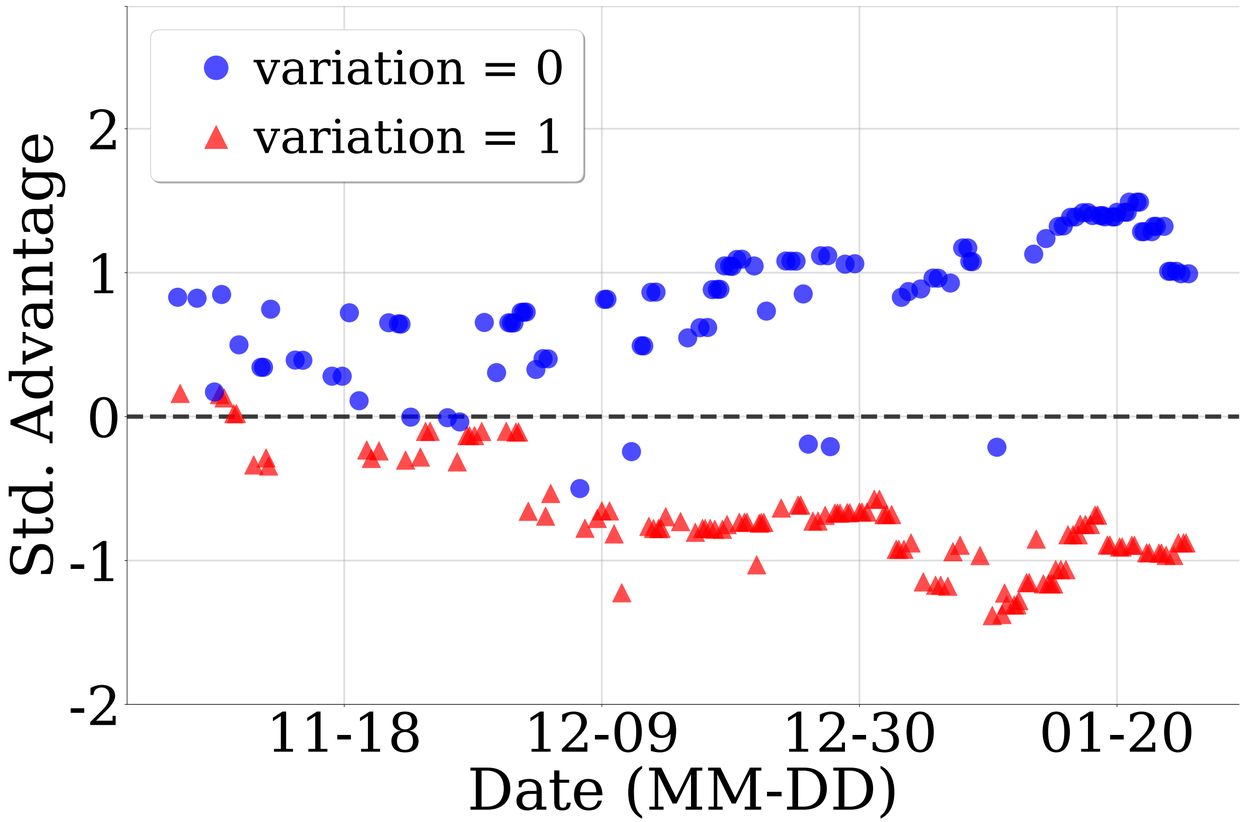
- We retain $\{S_t, \hat{\varepsilon}_t\}_{t=1}^{T}, \hat{\alpha}_T, \hat{\beta}_T$ for each user
    - Changing $\hat{\beta}_T$ changes the "ground truth" for a user
    - A generative model allows us to generate potential outcome for any action in the resampled/resimulated trajectory
    - We rerun the RL algorithm for each user
- For each set of resampled user trajectories we compute the interestingness score for each trajectory and count the number of interesting users

Ideas implicit in our framework

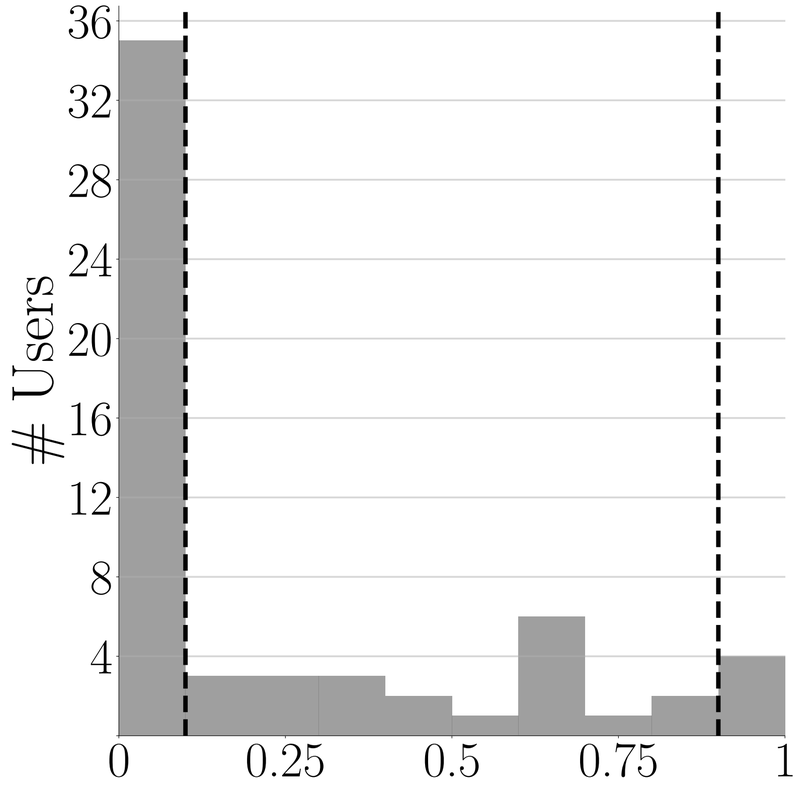
- To investigate a particular notion of interestingness, we require
  - specifying a generative model tailored to that interestingness
  - and having an understanding for the desired behavior for the RL algorithm under that generative model

Interestingness of type 2:
What is a good generative β to investigate these questions?

Is RL personalizing for user 2?



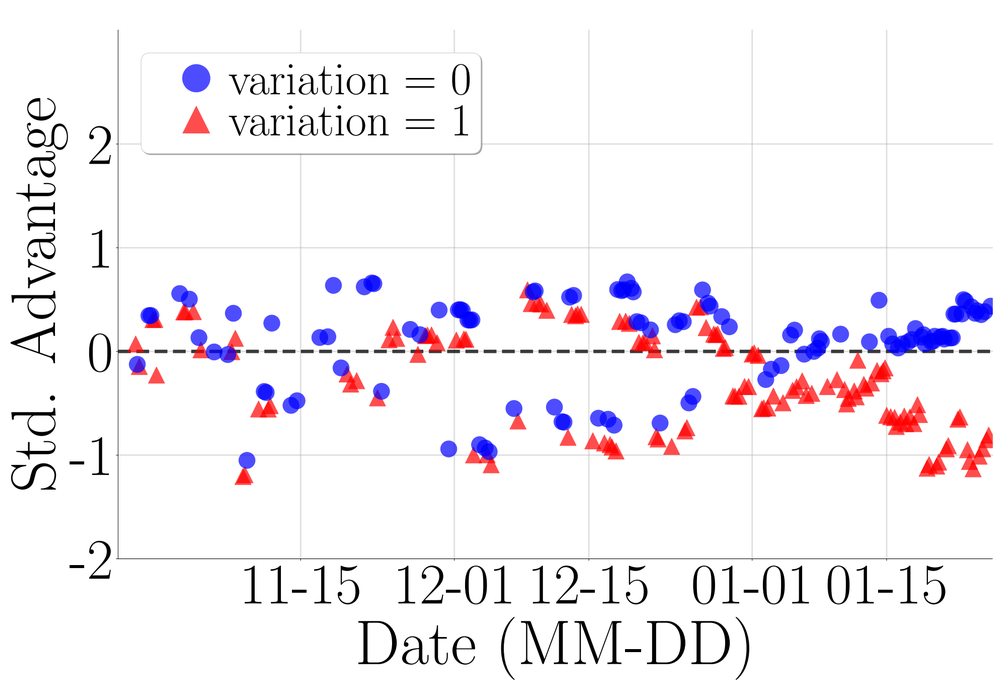
Did RL learn to personalize actions based on z = "variation" for a significant number of users?

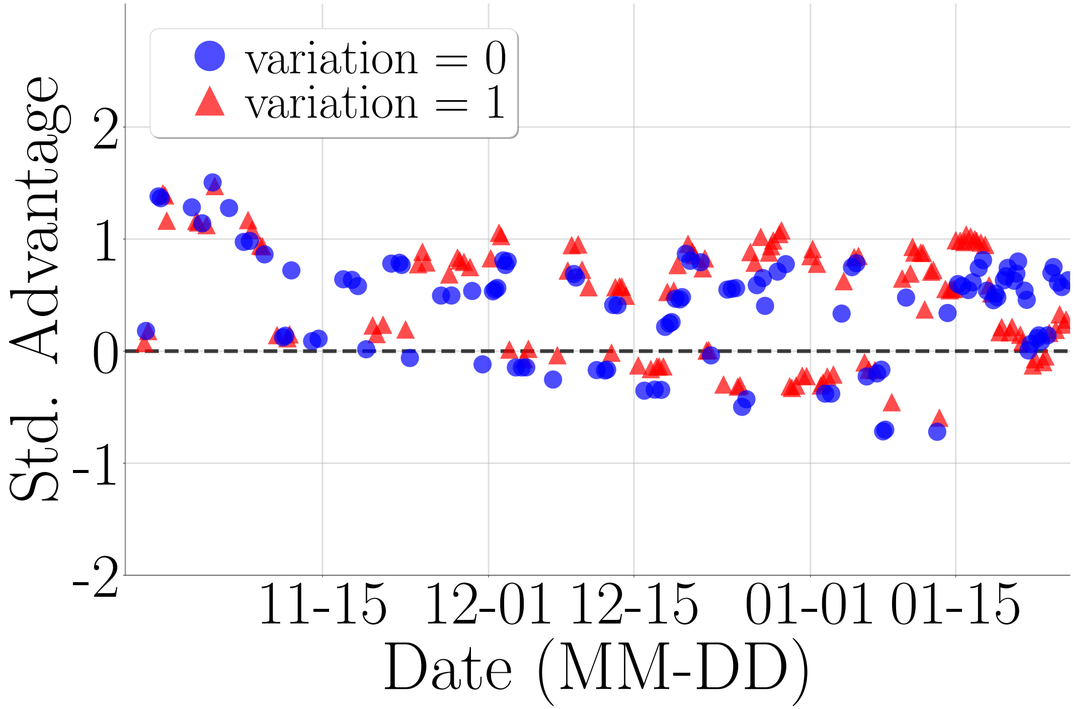


$$\text{Score\_int}_{2,z} = \frac{|\{t : \hat{\Delta}_t(S_{t,z=1}) > \hat{\Delta}_t(S_{t,z=0})\}|}{T}$$

User 2: Resampled trajectories with no interaction of advantage with variation ($\beta_{variation} = 0$)

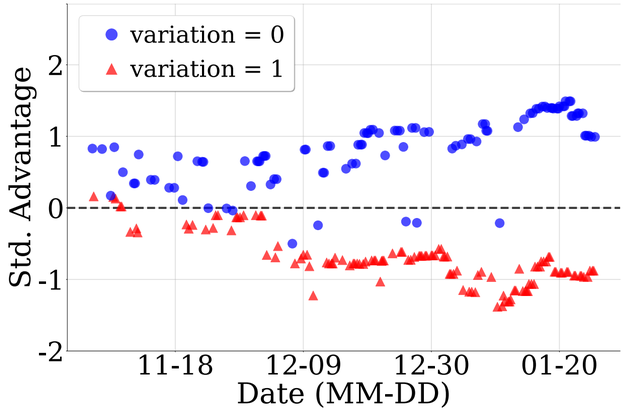
Other coefficients in the treatment model set equal to the value in the posterior mean



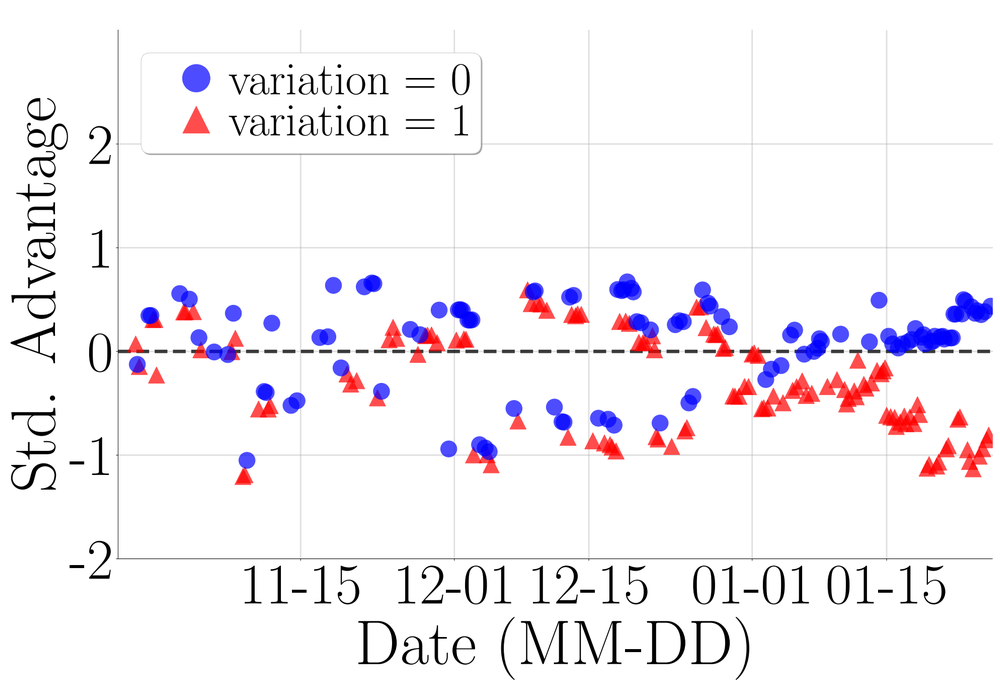


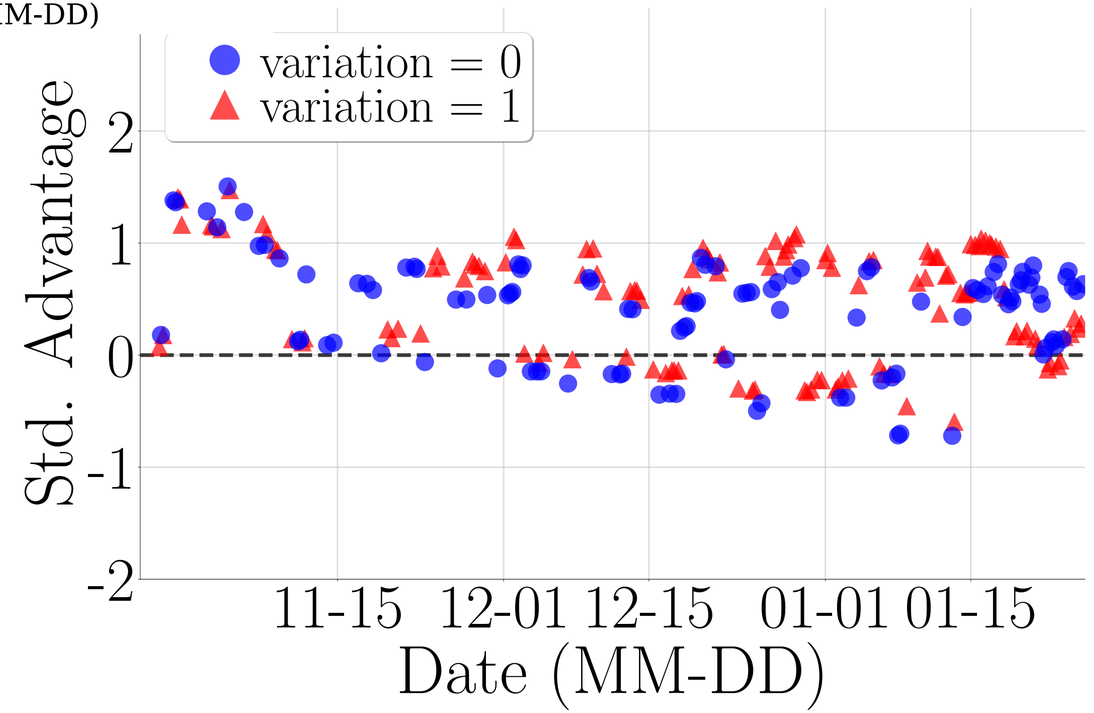
2 randomly chosen resampled trajectories out of 500

User 2: Resampled trajectories with no interaction of advantage with variation ($\beta_{variation} = 0$)



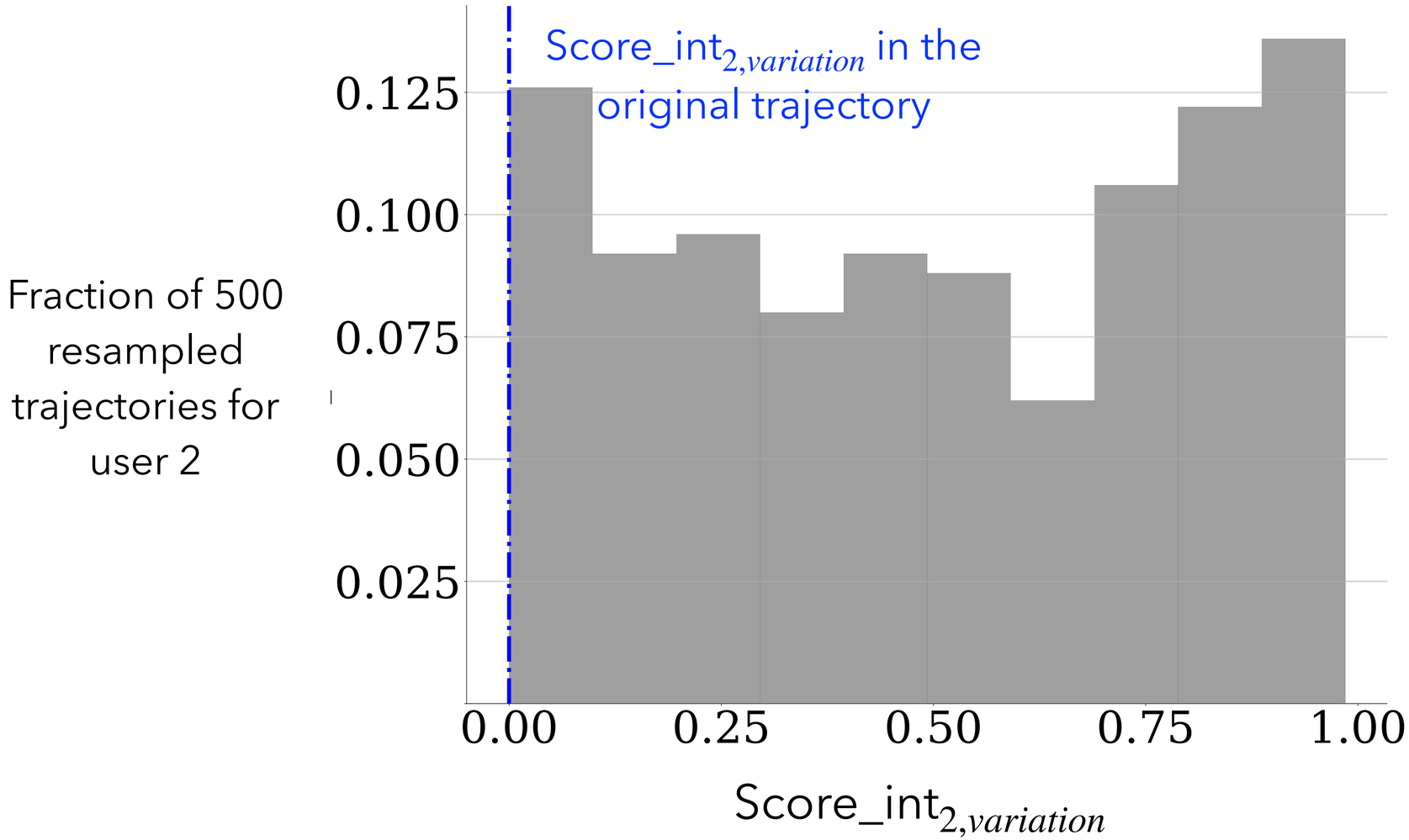
Observed/original trajectory



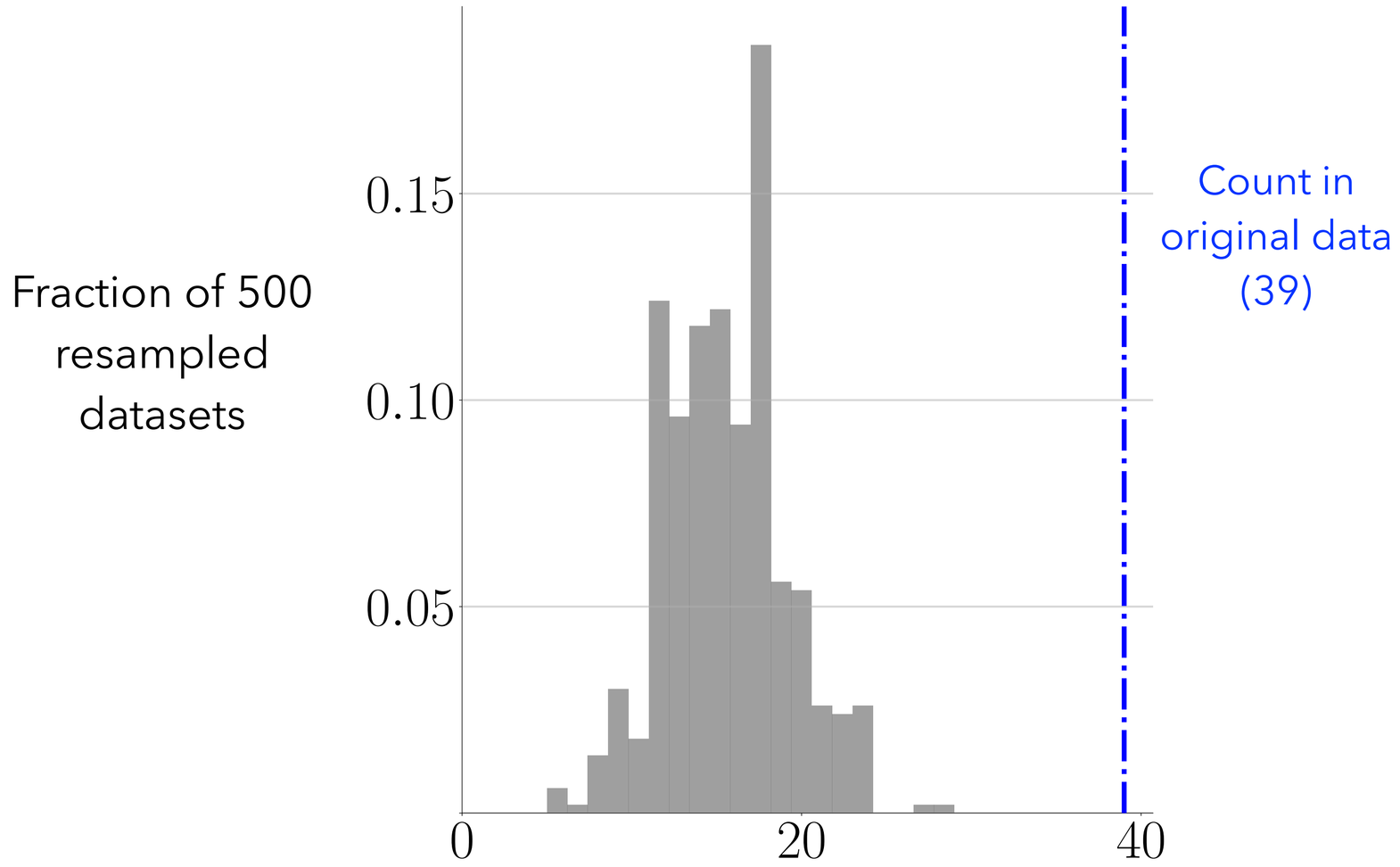


2 randomly chosen resampled trajectories out of 500

User 2: Resampled Score_int$_{2,variation}$ with no interaction of advantage with variation ($\beta_{variation} = 0$)



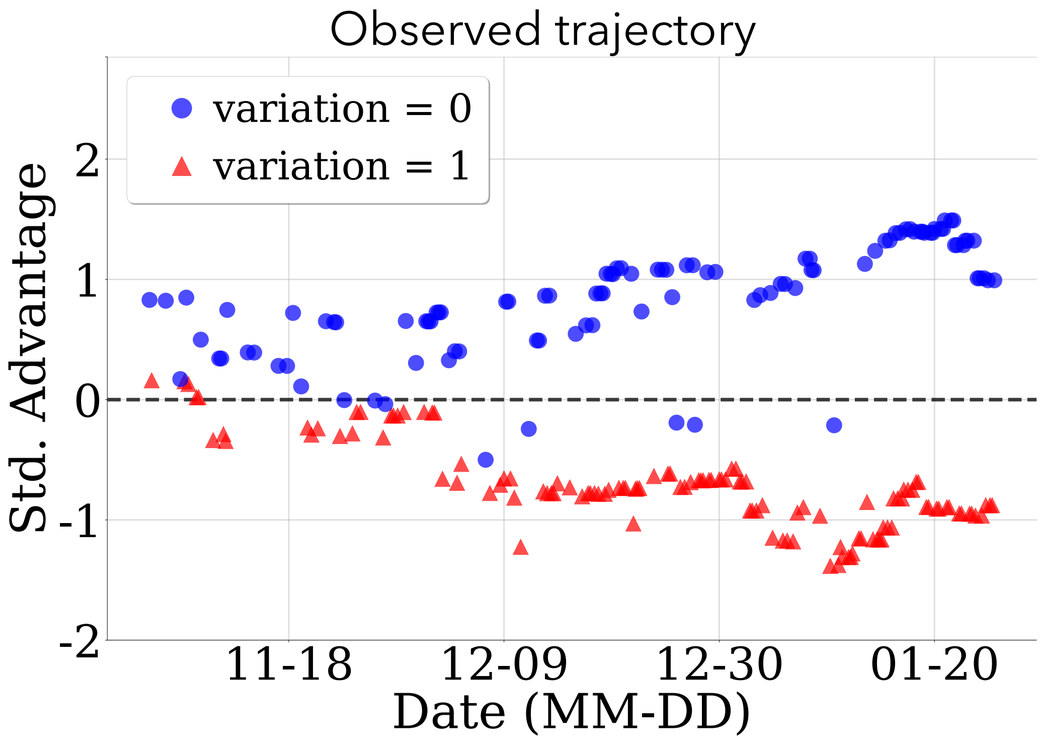
Number of interesting users of type 2 for variation across resampled datasets



- What other ways can we visualize if the algorithm is personalizing with respect to a feature?

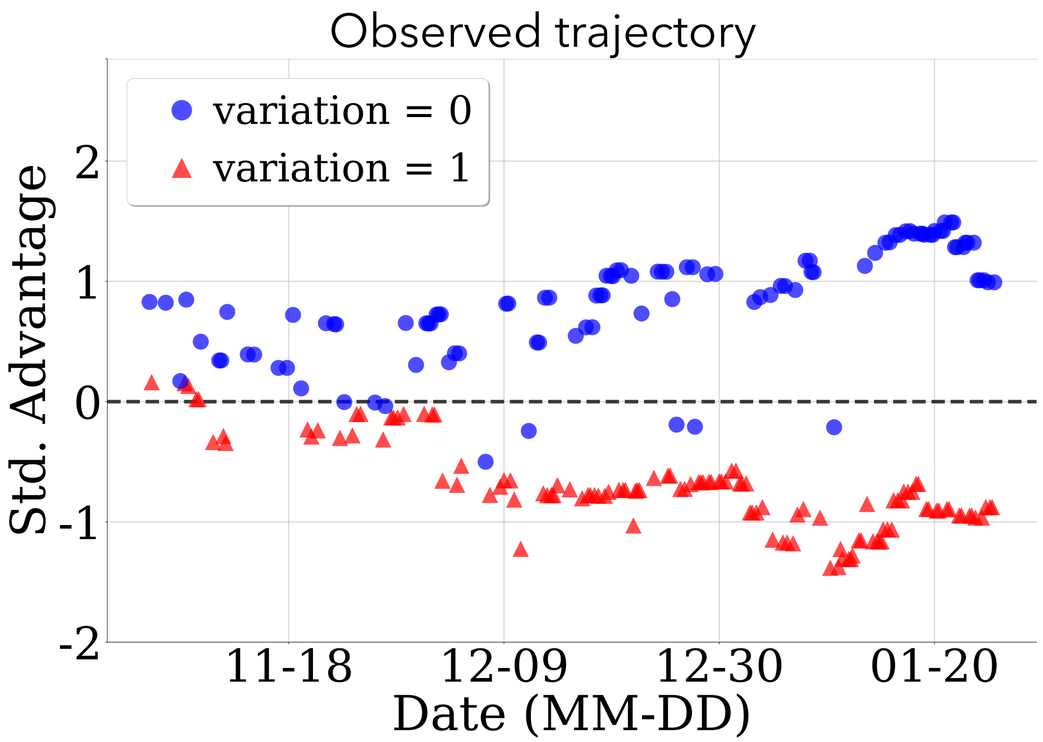
Alternate visualization of resampled trajectories for user 2 (interesting of type 2 for variation): standardized $\hat{\beta}_{t,variation}$

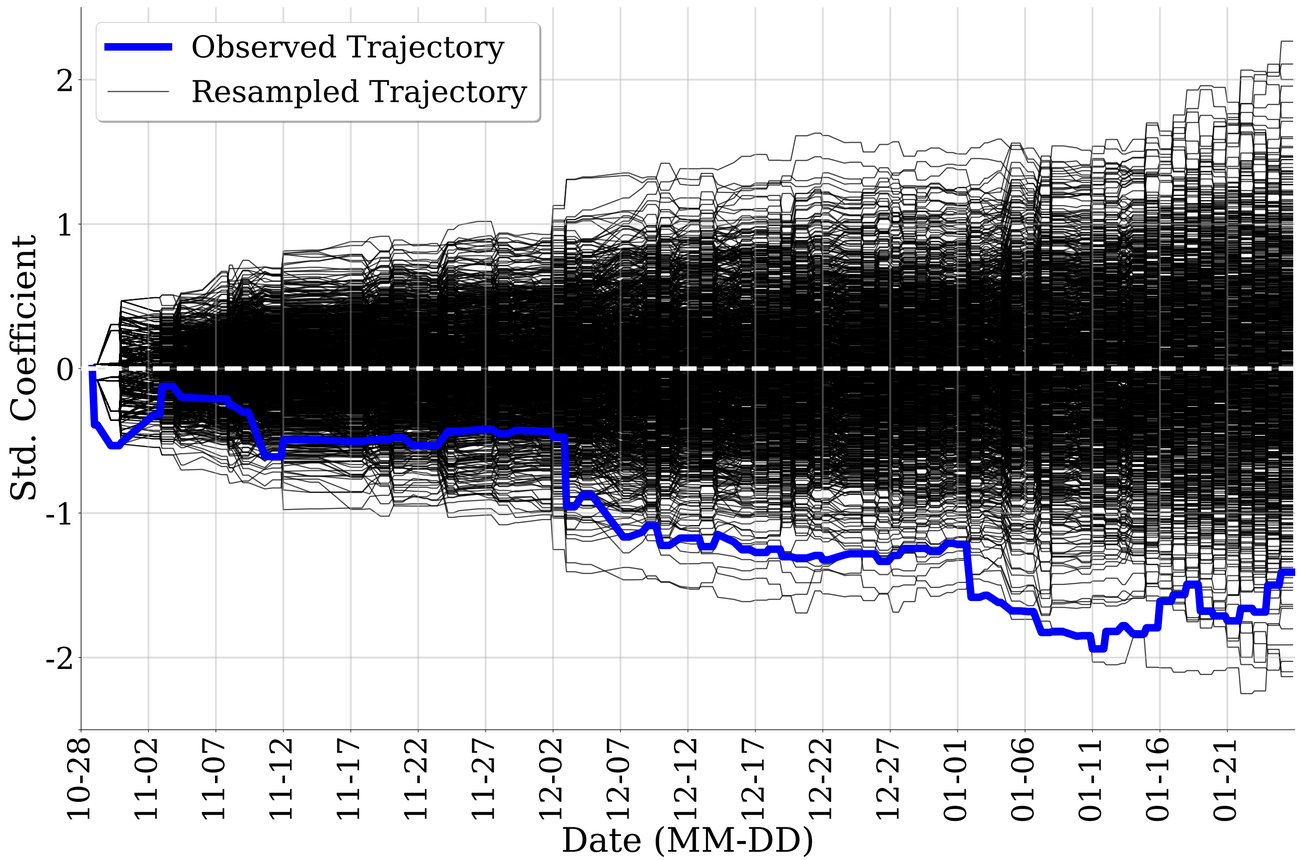
Variation {0, 1}: Whether the user's recent step counts have been highly variable (1) or not (0)



Alternate visualization of resampled trajectories for user 2 (interesting of type 2 for variation): standardized $\hat{\beta}_{t,variation}$

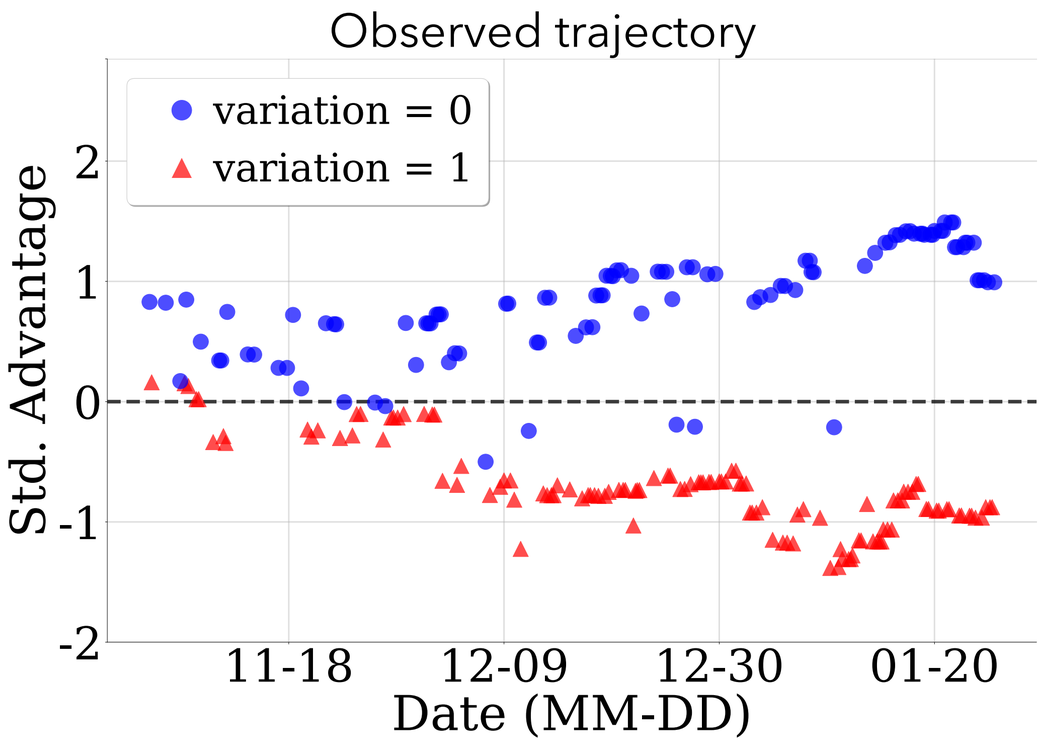
Variation {0, 1}: Whether the user's recent step counts have been highly variable (1) or not (0)

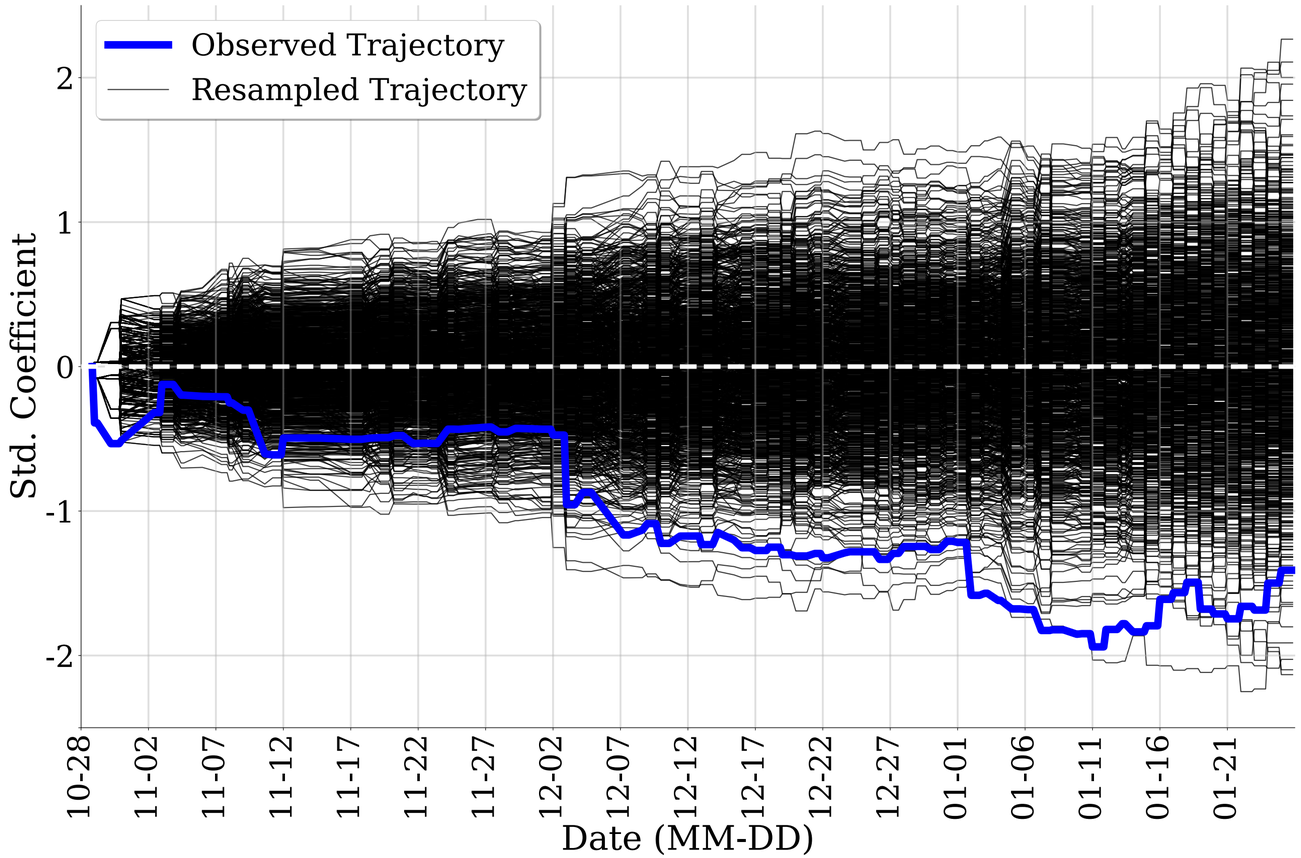




Alternate visualization of resampled trajectories for user 2 (interesting of type 2 for variation): standardized $\hat{\beta}_{t,variation}$

Variation {0, 1}: Whether the user's recent step counts have been highly variable (1) or not (0)

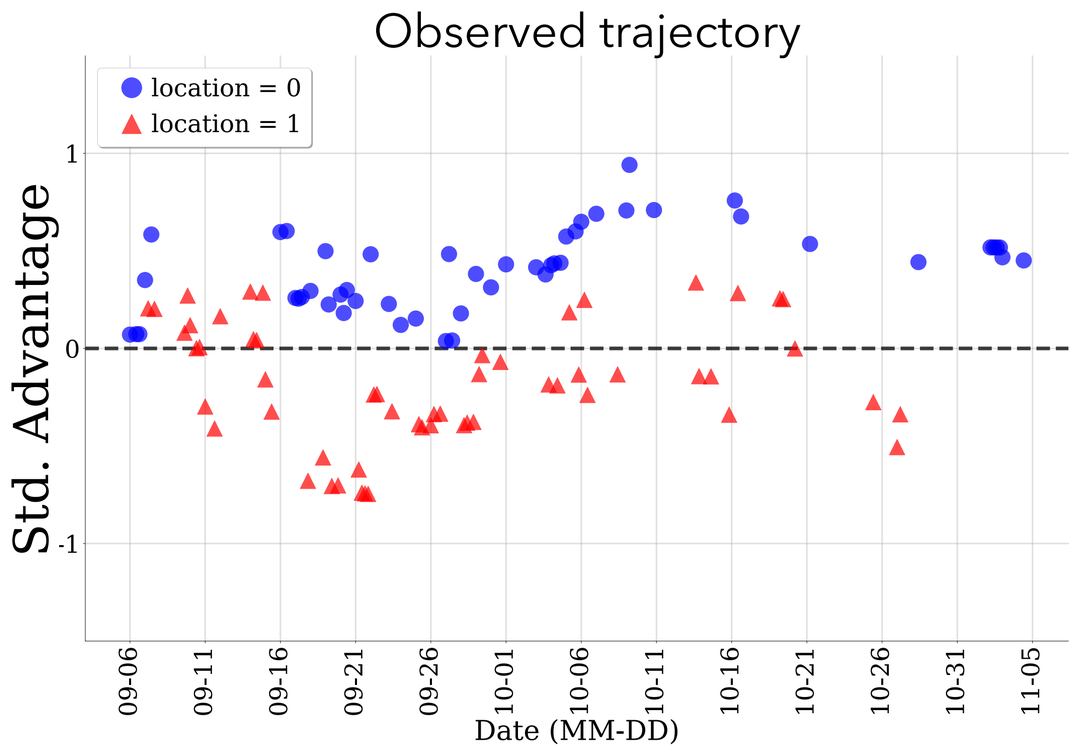


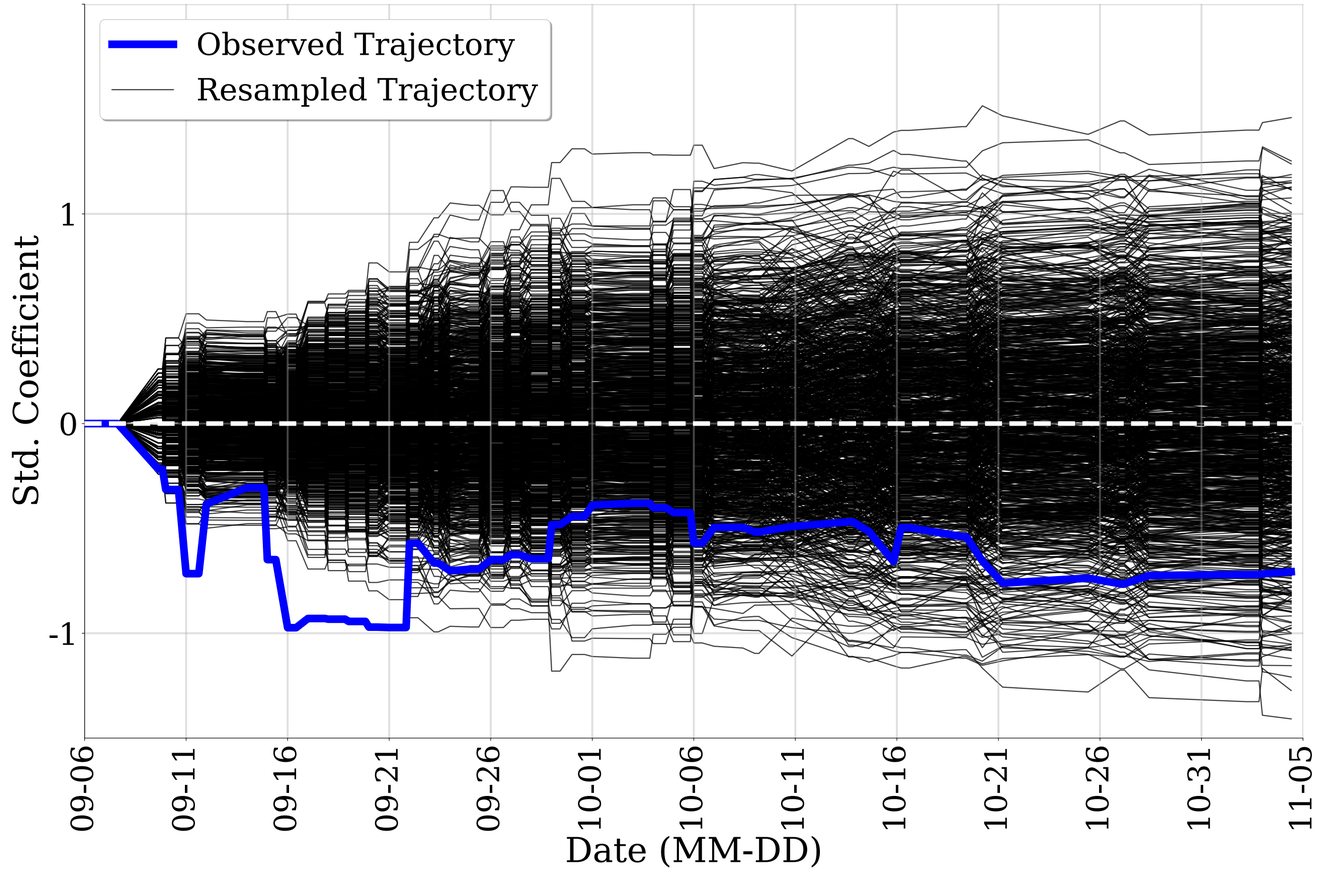


How would you characterize whether the blue curve is similar to the black curves or not?

Visualization of resampled trajectories for a user with interesting graph of type 2 for location (Right figure for standardized $\hat{\beta}_{t,location}$)

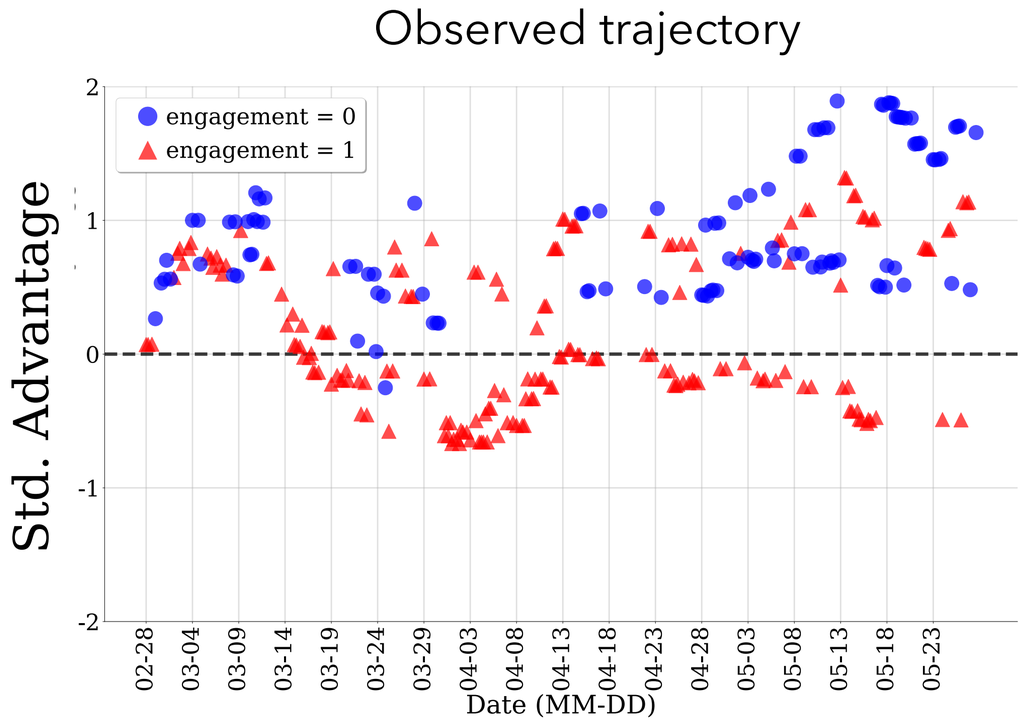
Location {0, 1}: Whether at home/work (0) or not (1)

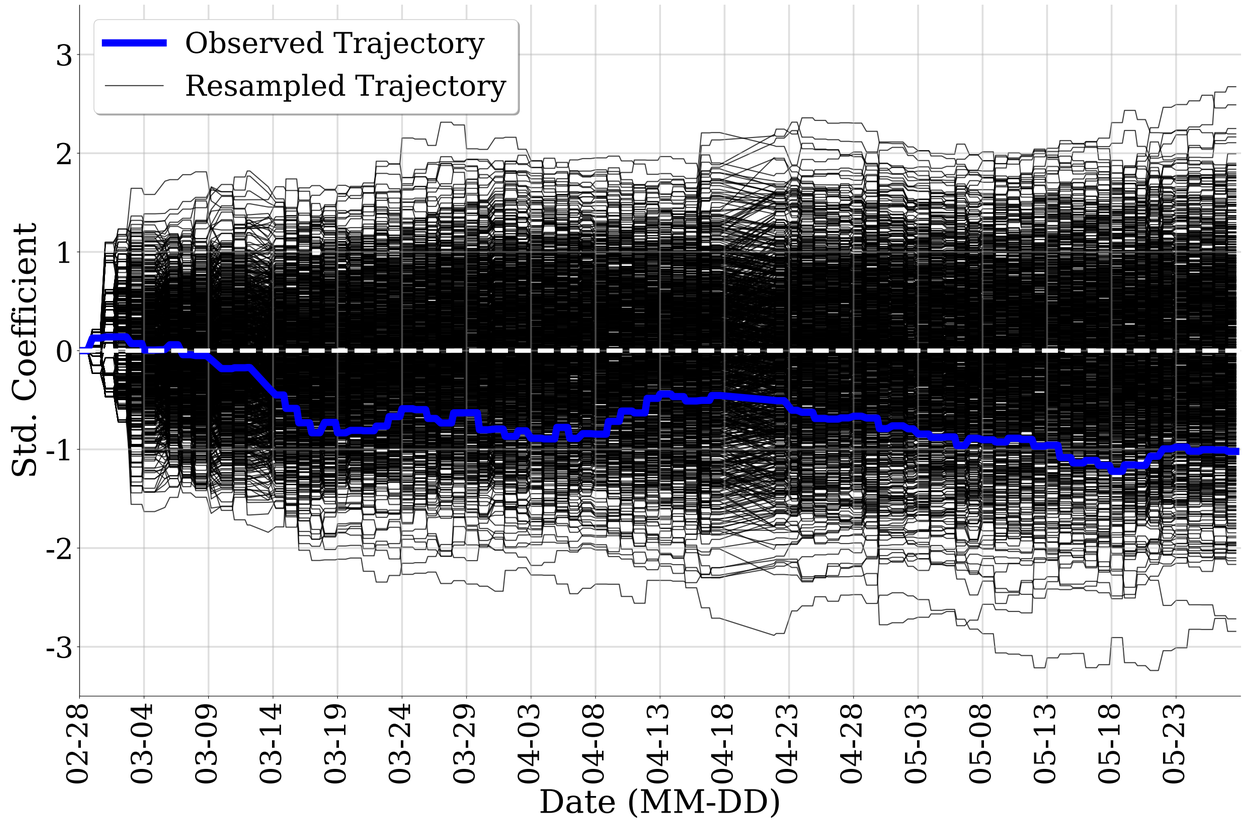




Visualization of resampled trajectories for a user with interesting graph of type 2 for engagement (Right figure for standardized $\hat{\beta}_{t,engagement}$)

Engagement {0, 1}: Whether the user has been interacting with the app (1) or not (0)



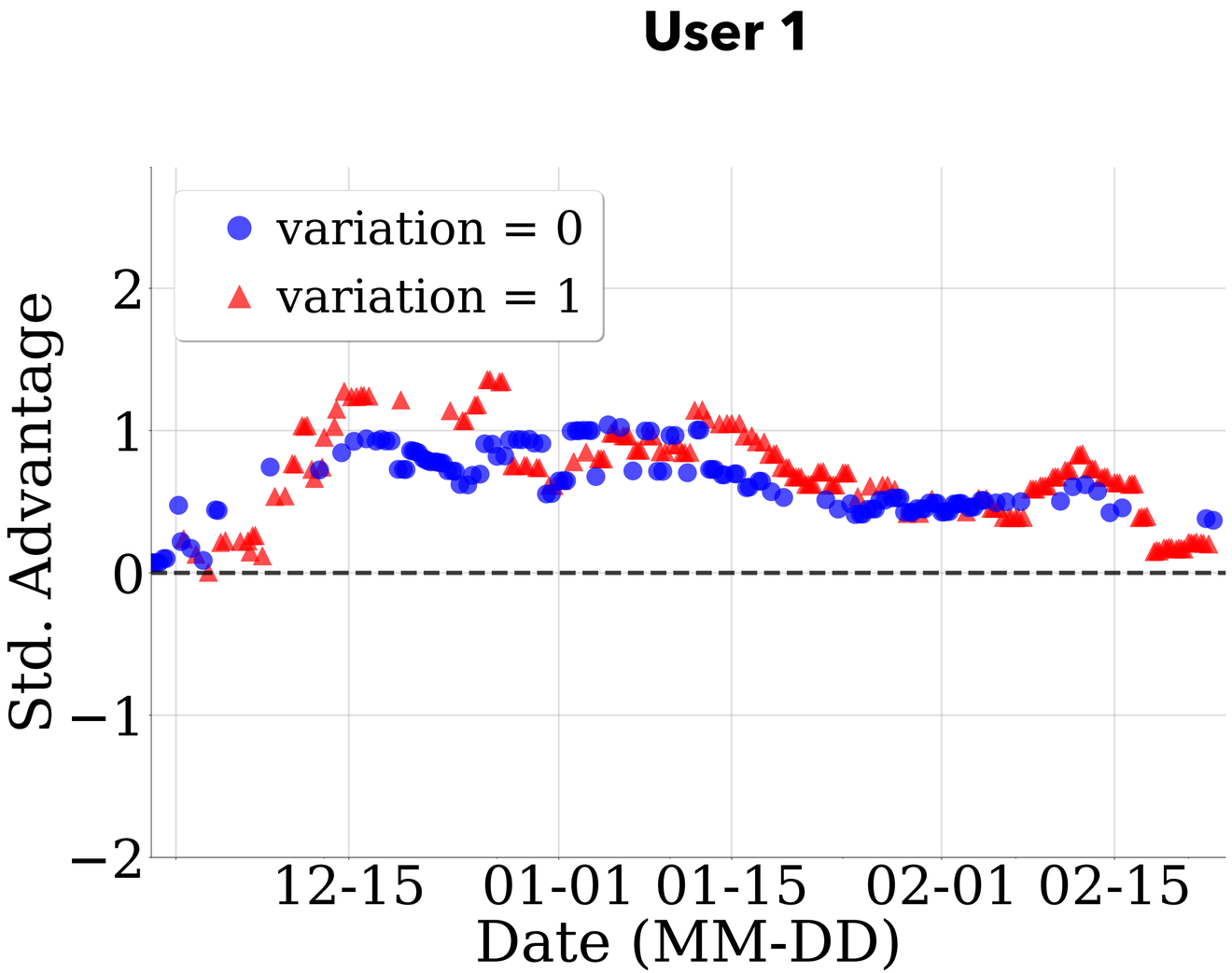


Discussion

- Does this approach provide some safeguarding from an RL algorithm that is learning to personalize slowly, or is faking personalization?
- What are possible future directions/extensions of this approach?
- How does this approach relate to the idea of stability: “If we perturb the data a little bit, the output of the algorithm should also not change too much?”

Appendix: Results for **interestingness of type 1**

Interestingness of type 1: Analysis

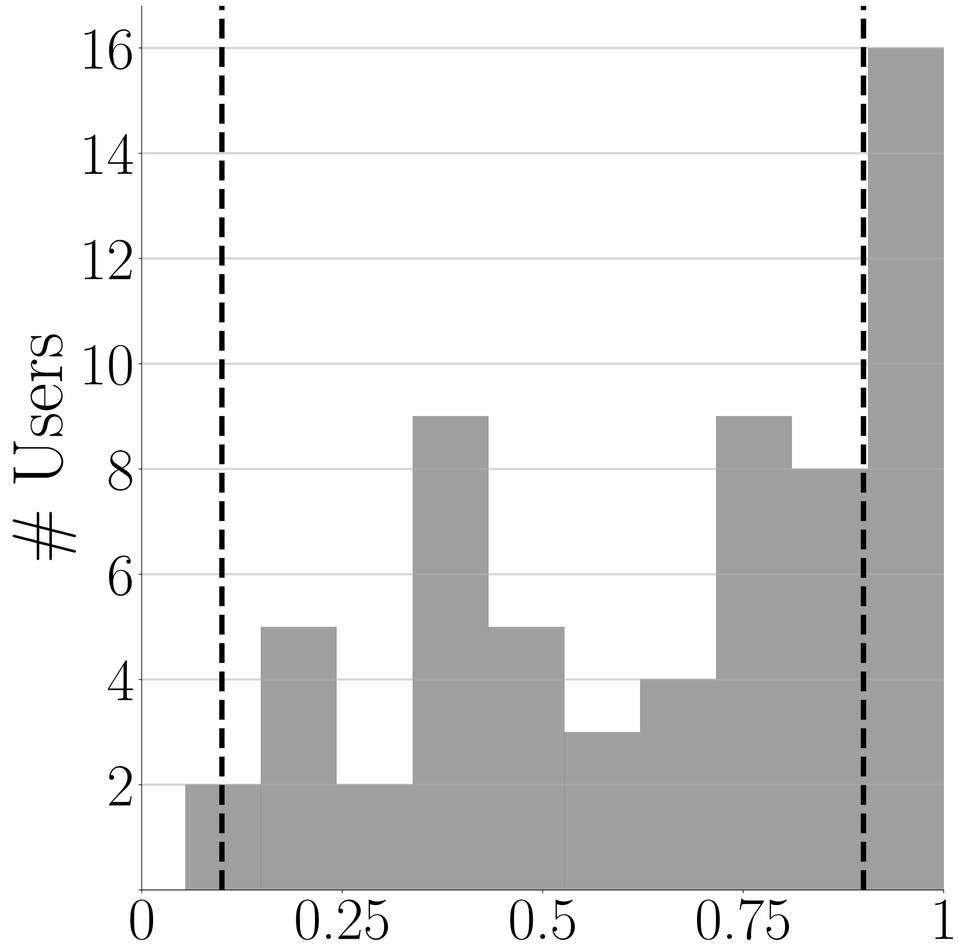


$$\text{Score\_int}_1 = \frac{|\{t : \hat{\Delta}_t(S_t) > 0\}|}{T}$$

Takes value 1 for this user!

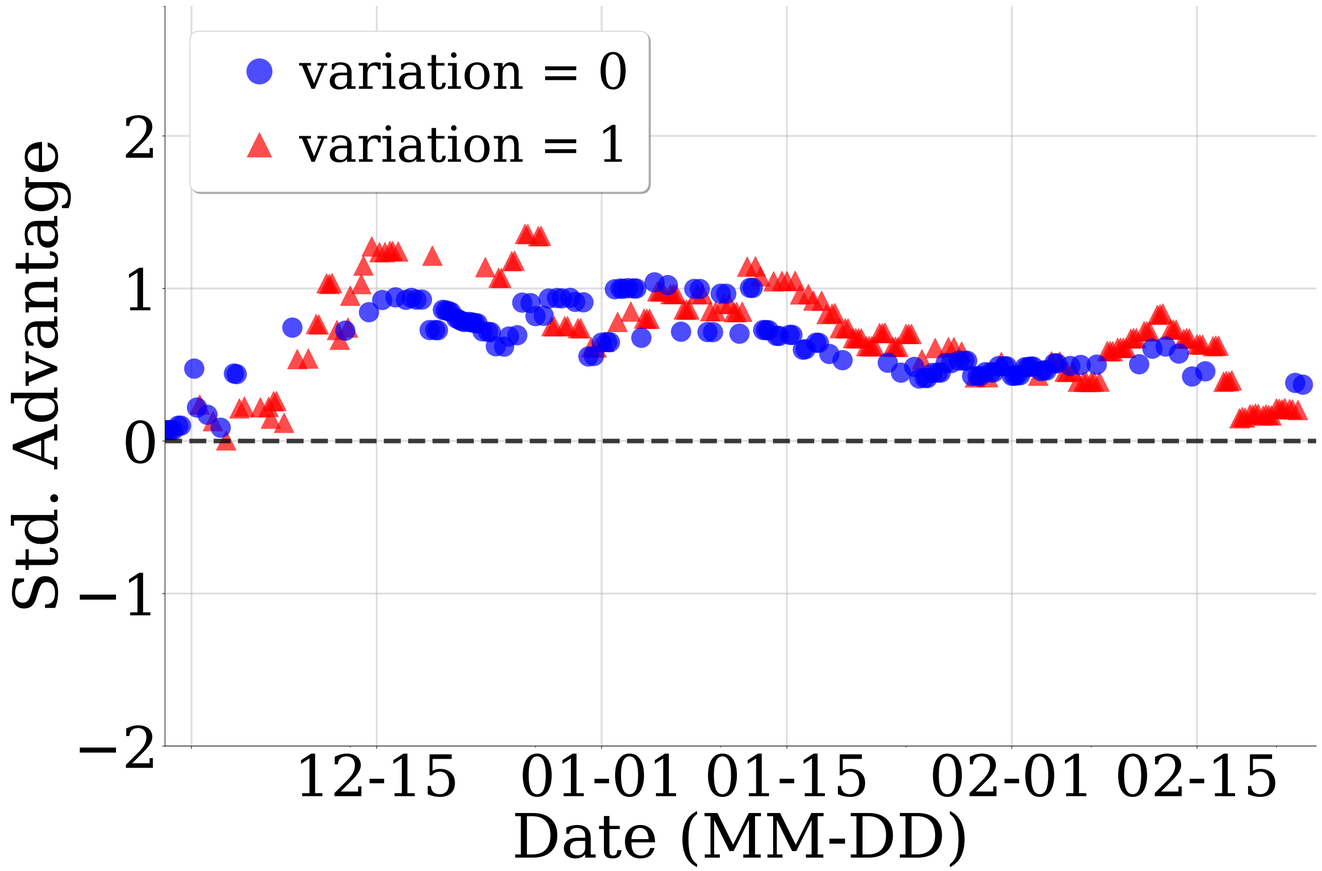
Interestingness of type 1: Values observed in the data

- Smoothen out the forecasts on a daily basis (at least two decision times per day)
- Filter out users with low availability
    - Why are we doing the previous two steps?
- Leaves 63 users and 18 out of these had scores $|\text{Score\_int}_1 - 0.5| \geq 0.4$

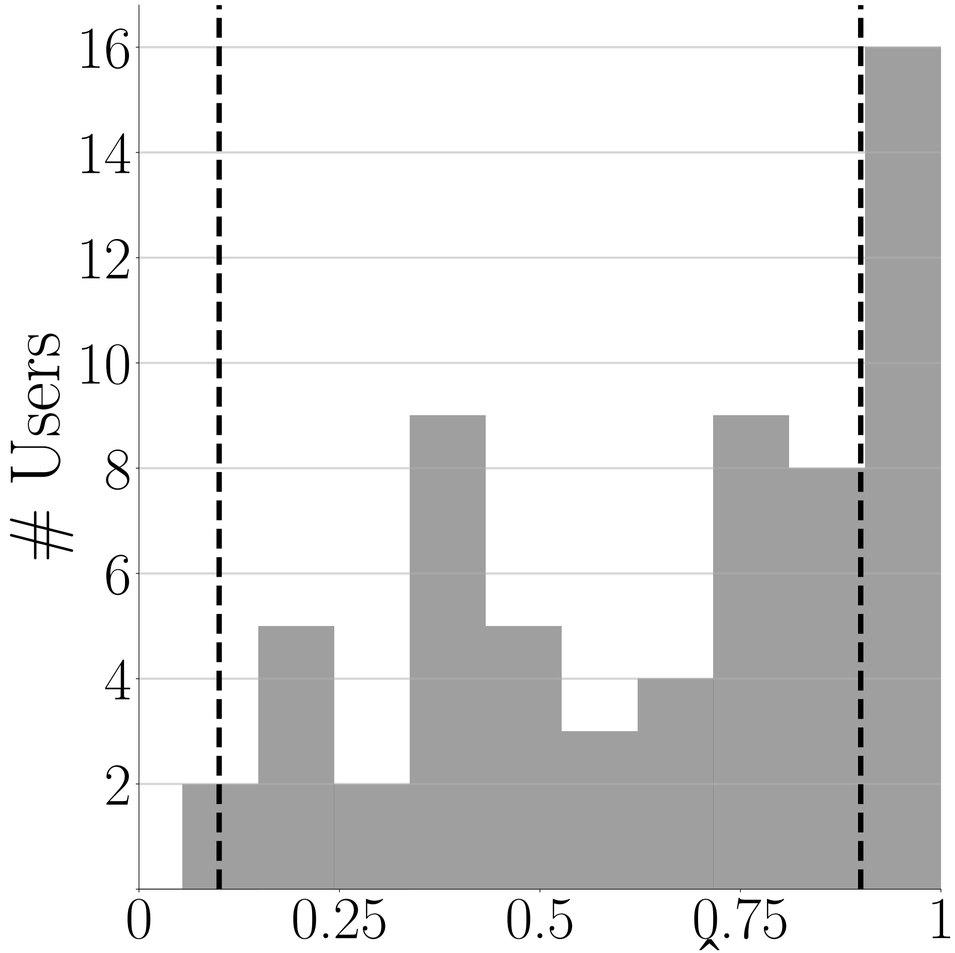


$$\text{Score\_int}_1 = \frac{|\{t : \hat{\Delta}_t(S_t) > 0\}|}{T}$$

Is RL personalizing for user 1?



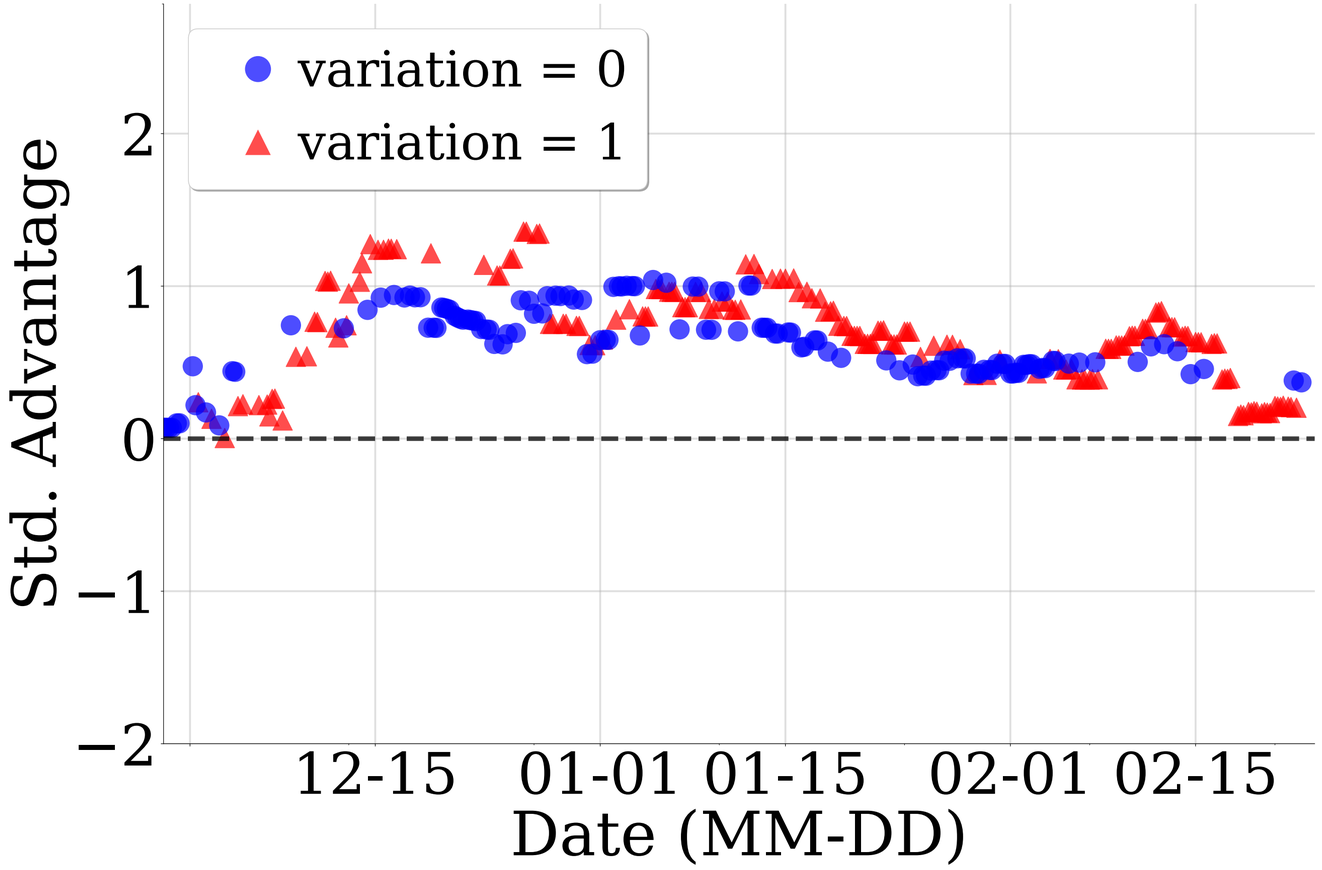
Did RL personalize for a significant number of users?



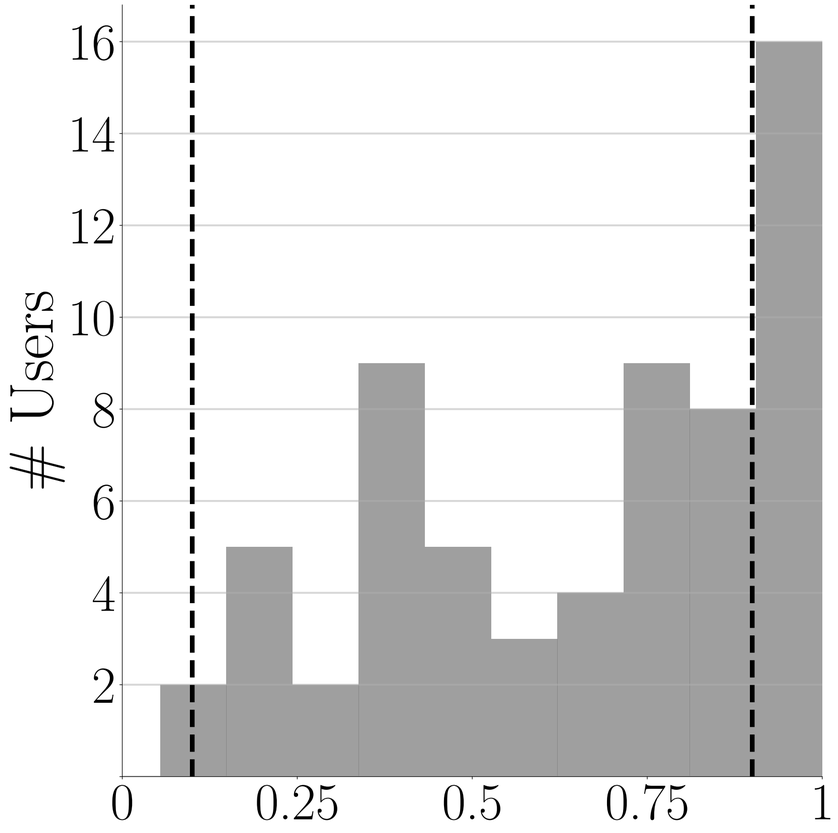
$$\text{Score\_int}_1 = \frac{|\{t : \hat{\Delta}_t(S_t) > 0\}|}{T}$$

Interestingness of type 1:
What is a good β for these questions?

Is RL personalizing for user 1?

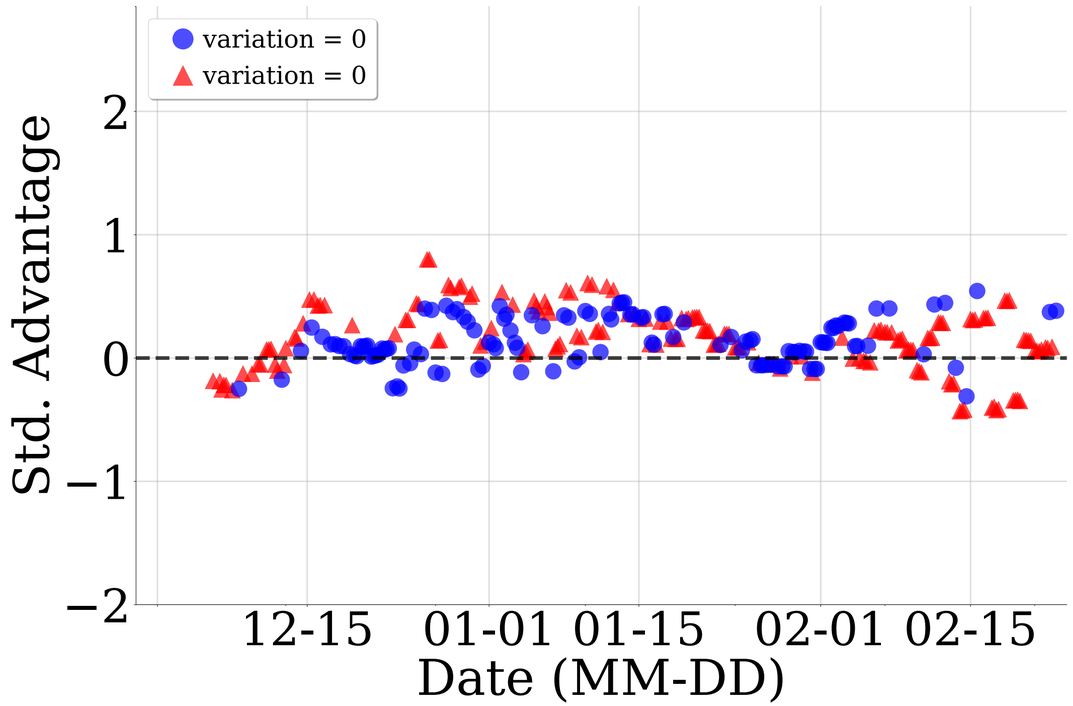


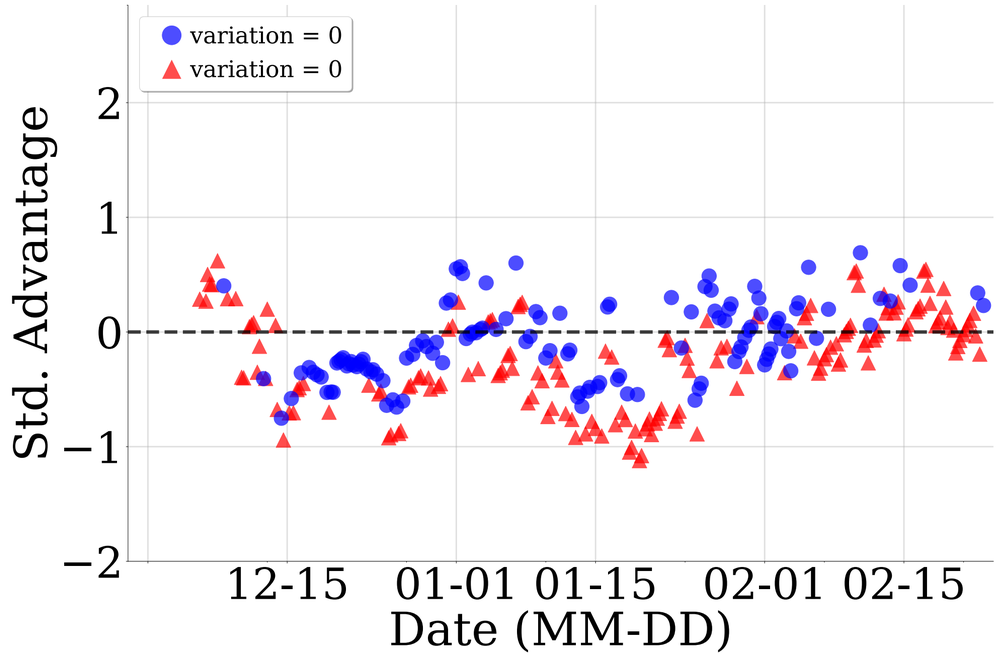
Did RL personalize for a significant number of users?



$$\text{Score\_int}_1 = \frac{|\hat{\Delta}_t(S_t) > 0|}{T}$$

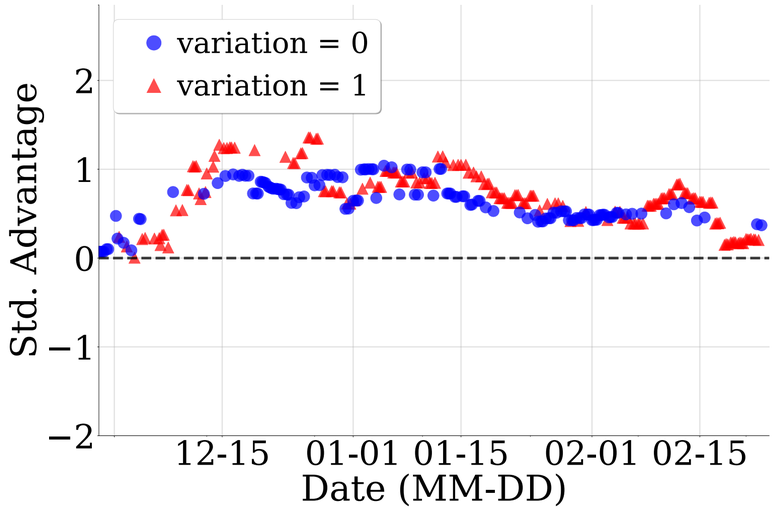
User 1: Resampled trajectories with no advantage ($\beta = 0$)



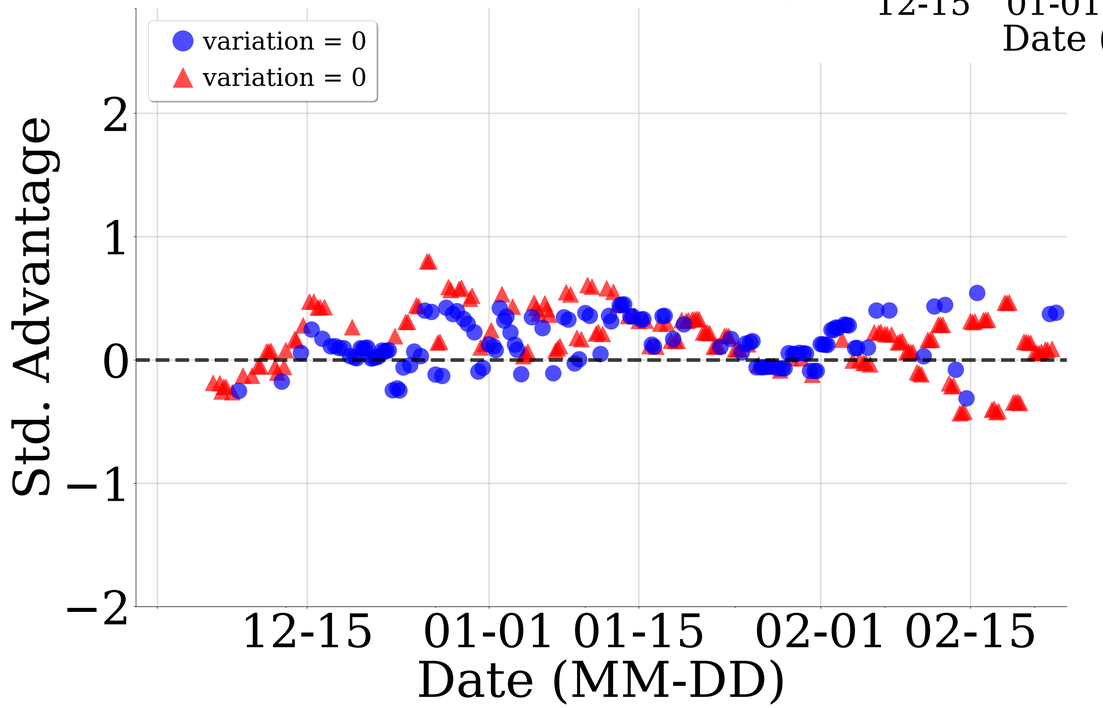


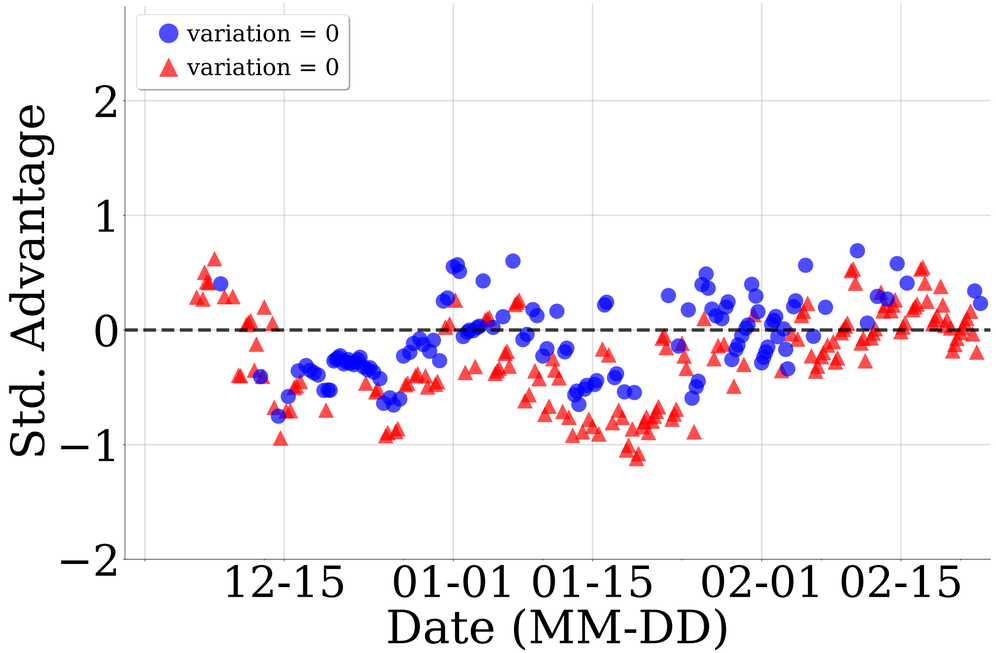
2 randomly chosen resampled trajectories out of 500

User 1: Resampled trajectories with no advantage ($\beta = 0$)



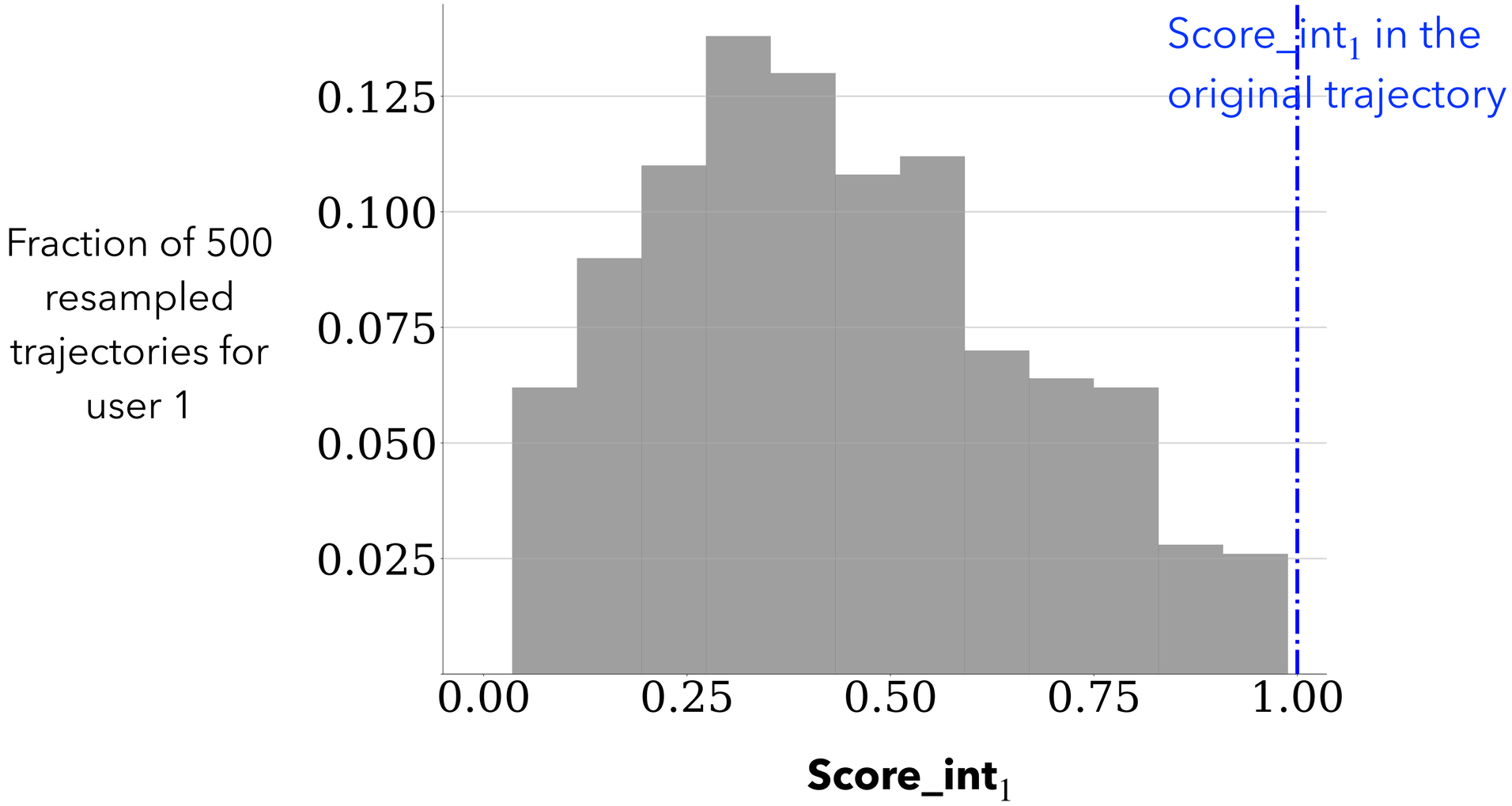
Observed/original trajectory



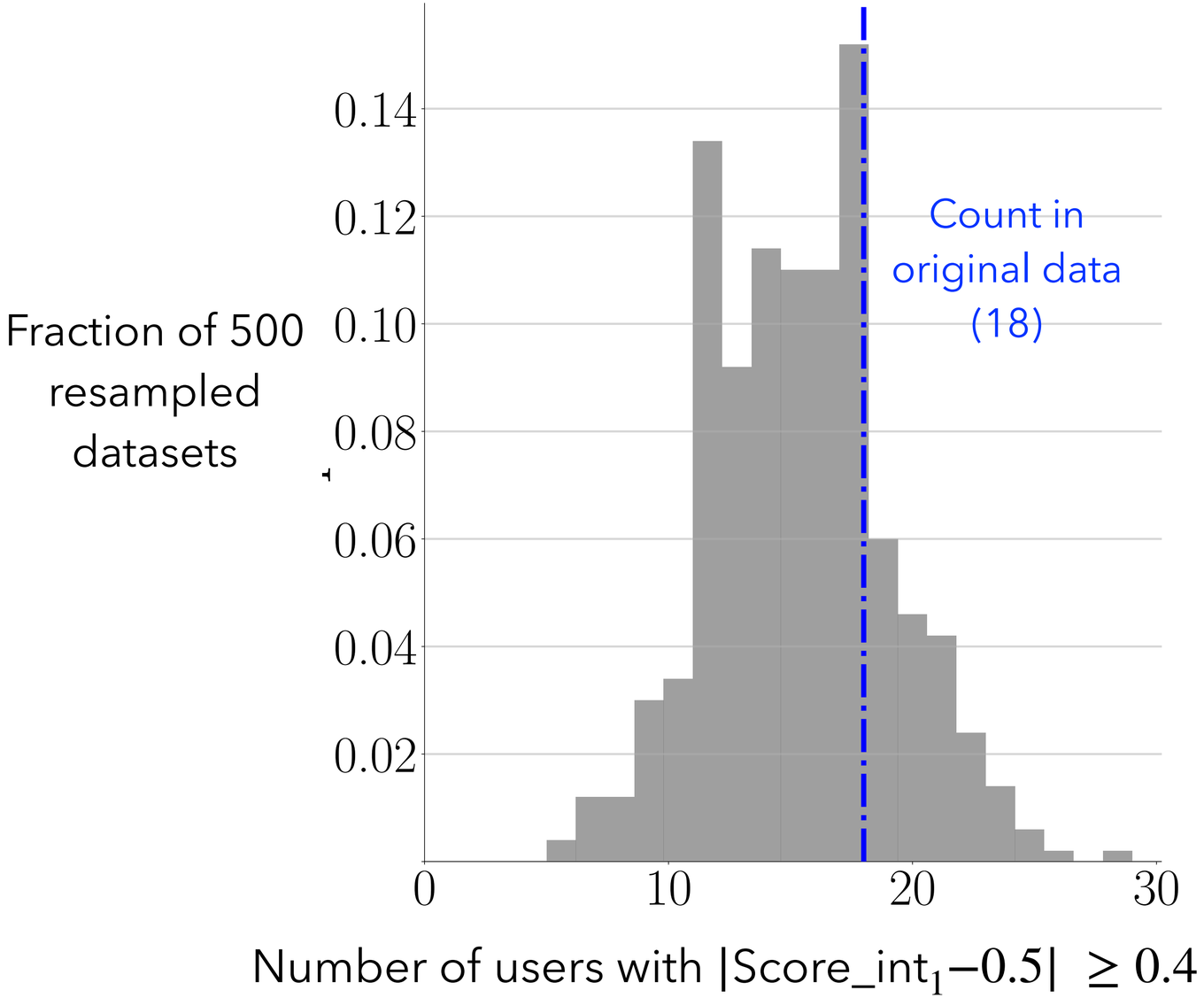


2 randomly chosen resampled trajectories out of 500

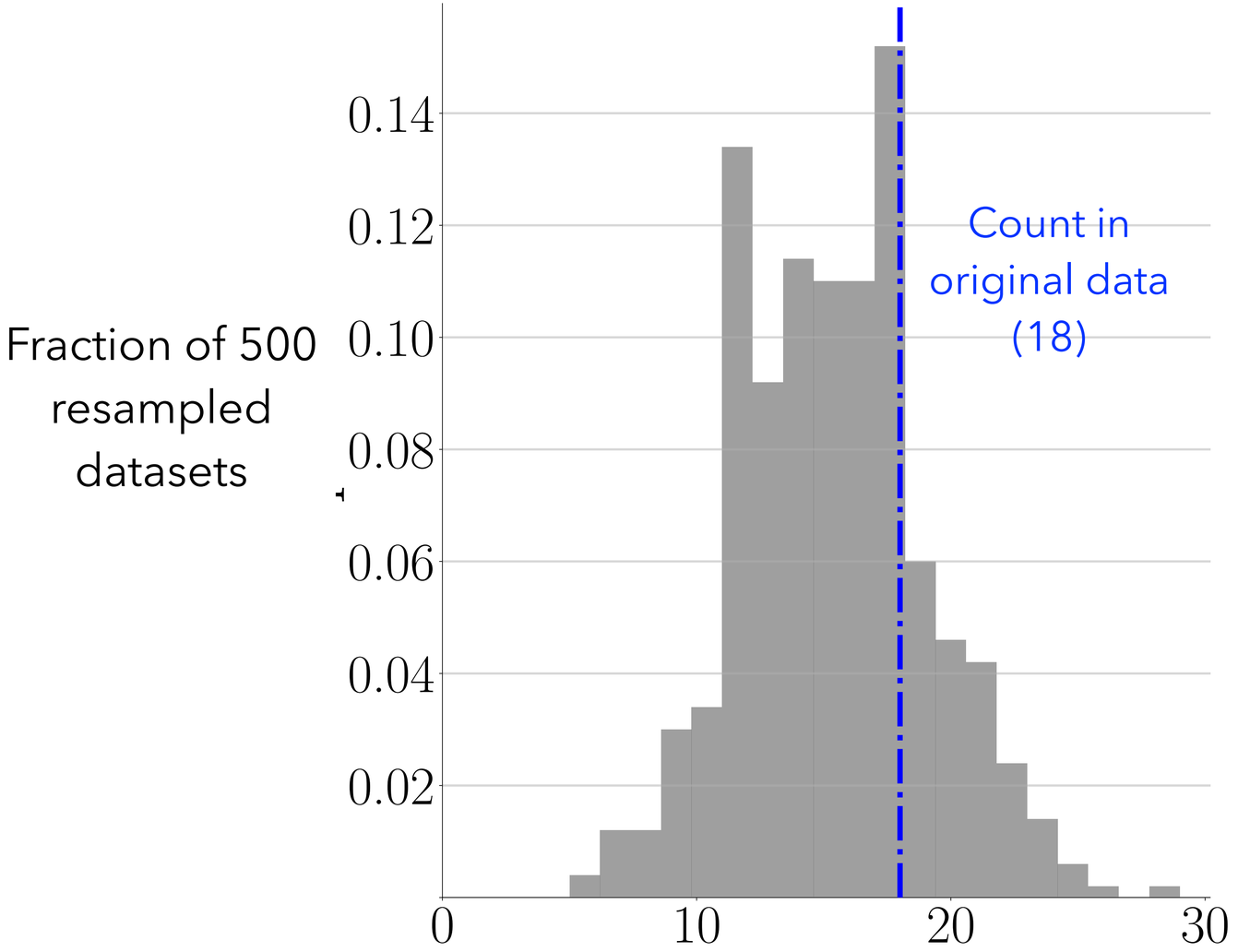
User 1: Resampled Score_int$_1$ with no advantage ($\beta = 0$)

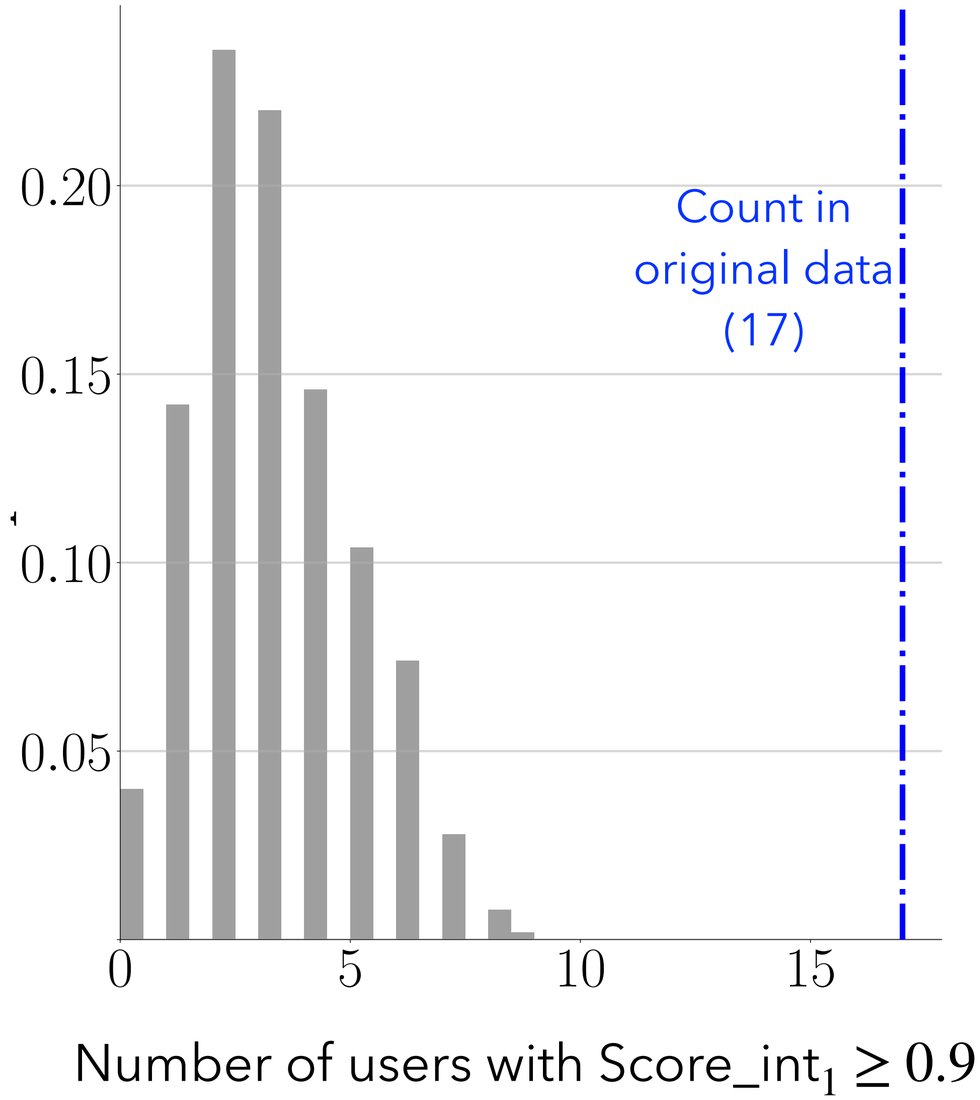


Number of interesting users of type 1 across resampled datasets



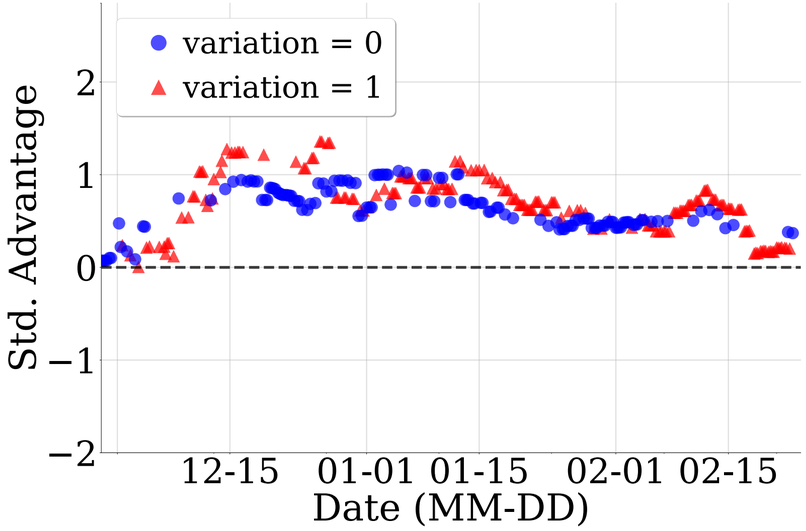
Number of interesting users of type 1 across resampled datasets



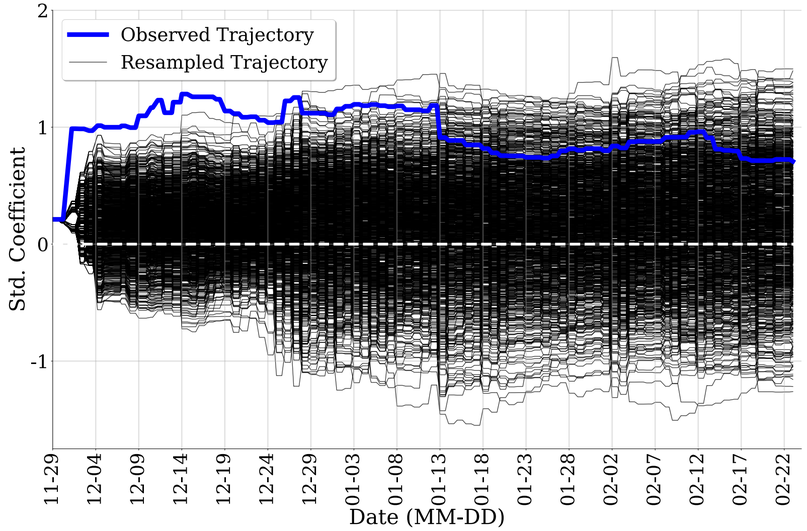


Visualization of user 1's trajectories for all features

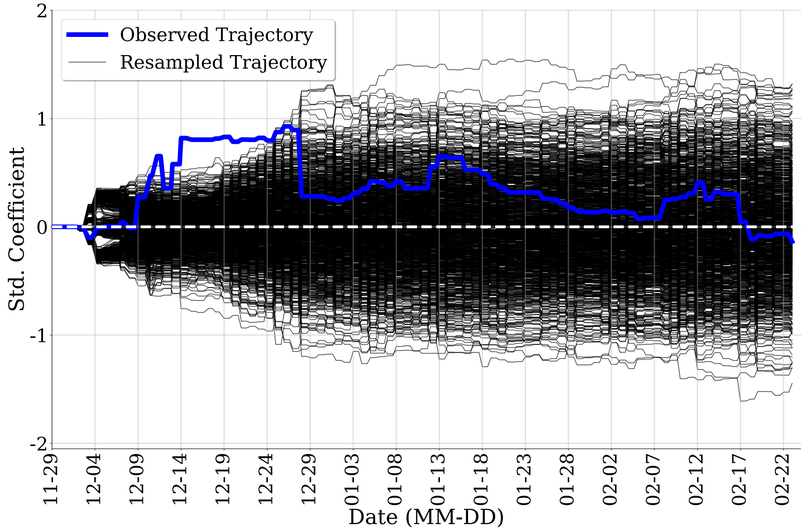
Observed trajectory



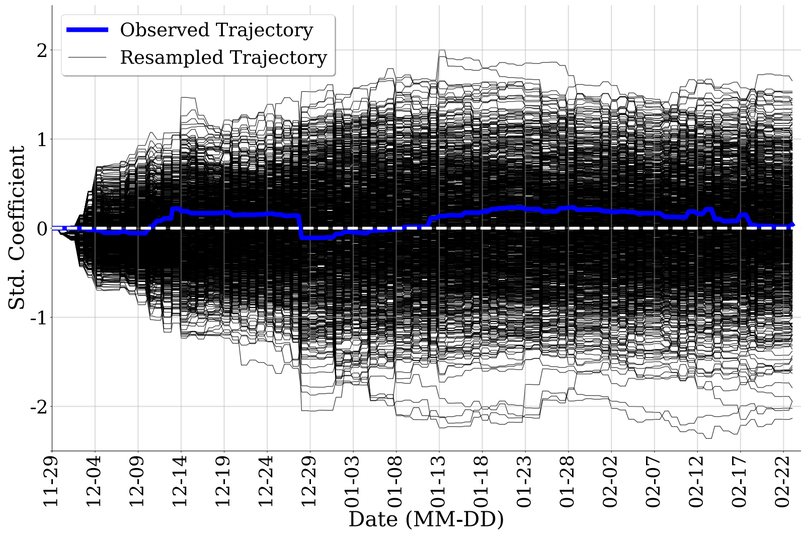
Intercept



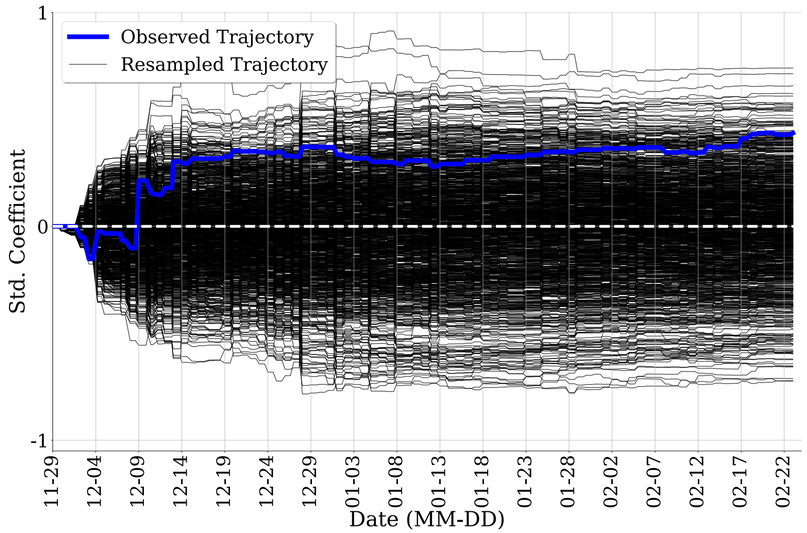
Variation



Engagement



Location



Dosage