User and Session Heterogeneity in Digital Experiments: A Framework for Analysis and Understanding

Sriram Somanchi[1], Ahmed Abbasi[1], David Dobolyi[1], Ken Kelley[1], and Ted Tao Yuan[2]
University of Notre Dame[1], eBay[2]

Background and Motivation

Digital experimentation platforms support data-driven decision-making at scale by allowing large quantities of online controlled experiments (OCEs) to be deployed simultaneously. OCEs are modeled after the classic randomized clinical trial concept, where one or multiple treatment settings are pitted against a status-quo or baseline control (Kohavi et al. 2020). With the value provided by OCEs widely known, demand for OCEs is increasing in digitally-focused organizational settings, which has led to some challenges. At a recent summit featuring digital experimentation leaders from a dozen major companies and academia, participants ranked the ability to effectively analyze OCE results as the number one overall concern (Gupta et al. 2019). Two analysis-related issues raised were: (1) user heterogeneity in treatment effects; (2) the impact of multiple concurrent experiments. Heterogeneity in treatment effects (#1) relates to potential for variance in treatment effects for sub-groups based on user characteristics such as prior behavior and engagement levels, as well as differences in devices, browsers, countries, and languages (Gupta et al. 2019; Chen et al. 2019). The impact of multiple concurrent experiments (#2) issue stems from the fact that in order to keep up with internal demand for more OCEs, the number of concurrent experiments has increased to the point where the same user is often in multiple experimental treatments concurrently (Gupta et al. 2019; Tang et al. 2010). User assignment in these scenarios is often intelligently managed through orthogonal random assignment of users to experiments that are unlikely to produce interactions. While highly effective in general, interactions are nonetheless still possible despite clever approaches to assignment (Kohavi et al. 2020; Xiong et al. 2020). To address these challenges, prior work has called for a two-pronged approach encompassing *detection* and *prevention* (Kohavi et al. 2013). Detecting effects for sub-groups, or multi-experiment interaction effects, is important especially when the effect sizes are drastically different or in a different direction from the average treatment effect (Gupta et al. 2019). Prevention entails making corrections to treatment policies enacted and/or the orthogonal experiment assignment mechanisms. Interestingly, as shown in Table 1, these challenges and the suggested mitigation strategies described in the OCE literature mirror the ongoing debate in the clinical trials space that digital experimentation is modeled on, namely the developing literature on *pragmatic trials* (Ford & Norrie 2016), which holds promise for OCE in the digital space.

Table 1: Parallels Between Select Challenges for Online Controlled Experiments and Pragmatic Clinical Trials

Challenge/Mitigation		**Online Controlled Experiments (OCEs)**	**Pragmatic Clinical Trials**
Challenge	User heterogeneity	Effect sizes can vary across user sub-groups defined by behavior or technology characteristics (Gupta et al. 2019; Chen et al. 2019).	Sub-populations can vary in effectiveness and reactions to treatments (Ford & Norrie 2016).
	Multiple experiments	Due to increased volume of experiments, sessions with exposure to multiple treatments might exhibit interaction effects (Kohavi et al. 2013; Chen et al. 2019; Kohavi et al. 2020; Xiong et al. 2020).	In order to increase enrollments, many pragmatic trials reach out to existing trial participants. This can cause 10-20% of participants to be in multiple treatments (Cook et al. 2013; Myles et al. 2014).
Mitigation	Detection	Important to detect sub-groups or experiment interactions with major differences in effect sizes or directionality (Gupta et al. 2019).	Detecting effects for sub-populations or multiple treatment interactions in the trial phase could reduce post-market adverse events (Ford & Norrie 2016).
	Prevention	In some cases, detected differences might warrant inclusion in result reports or changes to orthogonal assignment mechanism (Kohavi et al. 2013).	For severe adverse events, certain sub-groups or treatment interactions may be added to trial exclusion protocols and/or warning labels (Harron et al. 2012).

Proposed HTE Framework for Analyzing User-Session Level Digital Experiments

We propose a framework geared towards tackling these two challenges. *The primary intuition for our proposed framework is that the two aforementioned challenges are not tangential to one another – rather, both manifest at the user-session level as the amalgamation of user- and session-based heterogeneity.*

Based on this intuition, we propose a heterogeneous treatment effect (HTE) framework for analyzing user-session level digital experiments (Figure 1). The right side of Figure 1 illustrates our inputs into the HTE models. Given a focal experiment of interest A, with $i = 1, \dots, n$ user-sessions, the goal of an HTE model, with standard unconfoundedness assumption, is to estimate the treatment effect $\tau(x_i)$ given a user-session feature vector $X_i = x_i$, an outcome $Y_i \in \mathbb{R}$, and a treatment indicator $W_i \in \{0,1\}$. In our framework, X_i is comprised of user and session feature vectors U_i and S_i, respectively. U_i encompasses various user behavior characteristics such as prior engagement and/or expertise levels, etc. For S_i, we include a binary vector for a session's experiment treatment co-occurrence context. Given $j = 1, \dots, m$ experiments co-occurring with focal experiment A, $S_i \in [0,1]^m$, and with W_j indicating the control versus treatment setting for session i in experiment j, each $s_{ij} = 1$ only when $W_i, W_j > 0$. It is worth noting that while this formulation assumes a single treatment setting for each experiment, S_i can be extended to cases where $W_i, W_j \in \{0,1, \dots t\}$ by including all treatment-treatment s_{ij} combinations. Further, this vector could also be extended to include session temporal context indicators (based on timestamp information).

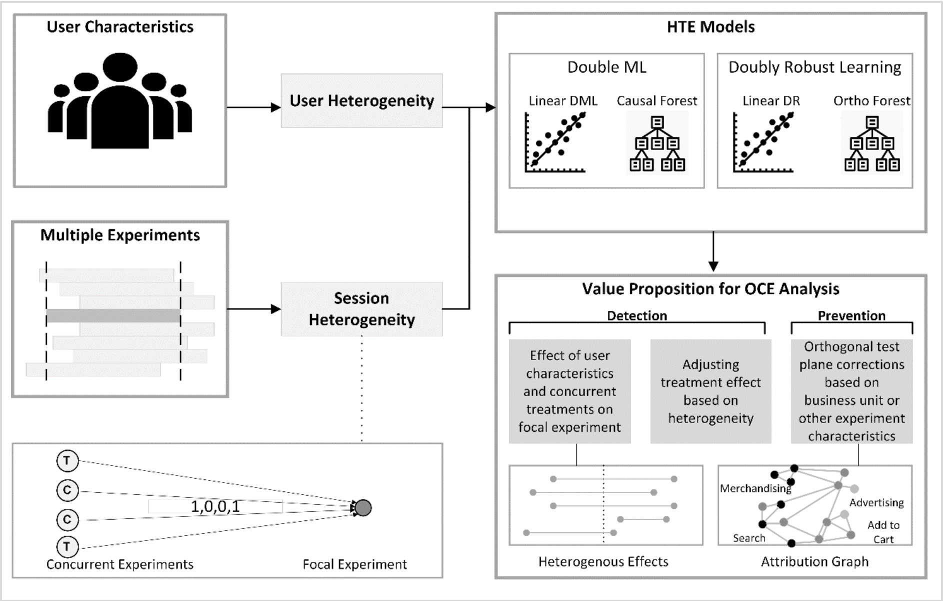

Figure 1: An HTE Framework for Analyzing User-Session Level Digital Experiments

Both double machine learning (DML) and doubly robust (DR) learning methods are used to model heterogeneous treatment effects, including linear DML, Causal Forest, Linear DR, and OrthoForest (Athey et al. 2019; Chernozhukov et al. 2018; Wager and Athey 2018). The tradeoff is that the linear models can more easily provide effect size estimates to convey the statistical results in a managerially meaningful way (Kelley and Preacher 2012) for each feature in X_i, but in the case of linear DML it is necessary to assume that the treatment effect on the outcome is also linear (Chernozhukov et al. 2018). The value proposition of our framework is threefold:

- Provides an HTE model-based estimate of the effect of user characteristics and concurrent treatments on each focal experiment.
- In cases where a focal experiment has many statistically significant interactions with other experiments, appropriately adjusting the effect size might be warranted. Similarly, in certain cases of user heterogeneity, the ideal policy/intervention might need to be tweaked.
- By combining results across multiple focal experiment sub-graphs, we can construct an *attribution graph* showing the interplay between concurrent experiments (e.g., how B affects A and vice-versa). This graph can 1) highlight important patterns related to concurrent experiments; 2) provide a mechanism for attributing the effects of experiments on one another; and 3) be rolled up to the

macro level (e.g., business unit-level attribution graph) to suggest possible tweaks to the orthogonal test plane setup if applicable.

Empirical Insights and Evaluation Plan

Our proposed framework is a method and a mechanism for generating empirical insights related to detection and prevention mitigation strategies for user heterogeneity and multiple experiment analysis challenges that manifest in OCEs. Using a large user-session testbed at a large online marketplace encompassing over 50 billion sessions related to 30 focal experiments and several hundred concurrent experiments spanning multiple product teams and business units, we will to present findings related to the following:

- Empirical Insights – We will present meta-level summary statistics on the pervasiveness of user heterogeneity in our context and multiple experiment interactions. Prior research suggests that these are challenges, yet it remains unclear if they are and to what extent will the interactions vary by the context of the situation. One would expect 5-10% of all user characteristics and/or experiment interactions to be significant purely by chance alone even if no heterogeneity existed – our work could be the first to empirically shed light on the size and scope of such heterogeneous effects.

- Evaluation – Prior work has noted the use of paired t-tests coupled with false discover rate methods as a way to identify potentially significant interactions between co-occurring experiment dyads (Kohavi et al. 2013; Kohavi et al. 2020; Strimmer 2008). We will evaluate whether the attribution graphs produced by our framework are capable of detecting major interactions (i.e., of significant effect size or where directionality changes) more efficiently than these existing "brute force" univariate approaches that rely on a large set of pair-wise comparisons that do not consider multiple features (i.e., X_i).

We believe the framework has important implications for OCE design-oriented research and practice in general, and provides specific value to our understanding of the online marketplace.

References:

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148-1178.

Chen, N., Liu, M., & Xu, Y. (2019). How A/B tests could go wrong: Automatic diagnosis of invalid online experiments. *ACM International Conference on Web Search and Data Mining (WSDM)*, 501-509.

Chernozhukov, V., Chetverikov, D., Demirer, M.,..., & Robins, J. (2018). *Double/debiased machine learning for treatment and structural parameters, The Econometrics Journal,* 21(1)*,* C1–C68.

Cook, D., McDonald, E., Smith, O., Zytaruk, N., ... & Meade, M. (2013). Co-enrollment of critically ill patients into multiple studies: patterns, predictors and consequences. *Critical Care*, *17*(1), 1-11.

Ford, I., & Norrie, J. (2016). Pragmatic trials. *New England Journal of Medicine*, *375*(5), 454-463.

Gupta, S., Kohavi, R., Tang, D., Xu, Y., Andersen, R., Bakshy, E., ... & Yashkov, I. (2019). Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter*, *21*(1), 20-35.

Harron, K., Lee, T., Ball, T., Mok, Q., Gamble, C., Macrae, D., ... & CATCH. team. (2012). Making co-enrolment feasible for randomised controlled trials in paediatric intensive care. *Plos One,* e41791.

Kelley, K., & Preacher, K. J. (2012). *On effect size*. Psychological Methods, 17(2), 137–152.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 1168-1176.

Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to A/B testing*. Cambridge University Press.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, *9*(1), 1-14.

Tang, D., Agarwal, A., O'Brien, D., & Meyer, M. (2010). Overlapping experiment infrastructure: More, better, faster experimentation. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 17-26.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228-1242.

Xiong, T., Wang, Y., & Zheng, S. (2020). *Orthogonal Traffic Assignment in Online Overlapping A/B Tests*, Tencent White Paper, EasyChair, No. 3110.