Fair Effect Attribution in Parallel Online Experiments

Alexander Buchholz
buchhola@amazon.com
Amazon Music ML
Berlin, Germany

Vito Bellini
vitob@amazon.com
Amazon Music ML
Berlin, Germany

Giuseppe Di Benedetto
bgiusep@amazon.com
Amazon Music ML
Berlin, Germany

Yannik Stein
syannik@amazon.com
Amazon Music ML
Berlin, Germany

Matteo Ruffini
ruffinim@amazon.com
Amazon Music ML
Berlin, Germany

Fabian Moerchen
moerchen@amazon.com
Amazon Music ML
Seattle, USA

ABSTRACT

A/B tests serve the purpose of reliably identifying the effect of changes introduced in online services. It is common for online platforms to run a large number of simultaneous experiments by splitting incoming user traffic randomly in treatment and control groups. Despite a perfect randomization between different groups, simultaneous experiments can interact with each other and create a negative impact on average population outcomes such as engagement metrics. These are measured globally and monitored to protect overall user experience. Therefore, it is crucial to measure these interaction effects and attribute their overall impact in a fair way to the respective experimenters. We suggest an approach to measure and disentangle the effect of simultaneous experiments by providing a cost sharing approach based on Shapley values. We also provide a counterfactual perspective, that predicts shared impact based on conditional average treatment effects making use of causal inference techniques. We illustrate our approach in real world and synthetic data experiments.

CCS CONCEPTS

• **General and reference** → **Experimentation**; *Measurement*; • **Theory of computation** → Solution concepts in game theory.

KEYWORDS

Cost Sharing, Causal Inference, Online Experiments, Shapley Values

ACM Reference Format:
Alexander Buchholz, Vito Bellini, Giuseppe Di Benedetto, Yannik Stein, Matteo Ruffini, and Fabian Moerchen. 2022. Fair Effect Attribution in Parallel Online Experiments. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3487553.3524211

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524211

1 INTRODUCTION

Randomized experiments like A/B tests [42], also known as randomized control trials (RCT), identify the effect of a treatment on a population. They are omnipresent on online platforms [21, 45]. Analyzing their outcomes allows to measure individual and interaction effects [33] for a large set of treatments in situations where treatments are perfectly randomized or potentially confounded [7]. However, massive simultaneous experimentation can have a cost for overall engagement metrics that the platform provider monitors. We provide an approach for measuring and disentangling the impact of parallel experiments and attributing their respective cost.

If several experiments are run on an online platform (for example, promoting new music content or showing several ads on a search engine), the overall user engagement measured across all experiments might decrease as users get overwhelmed. As an illustrative engagement metric we use throughout the click-through-rate (CTR). In this situation it is straightforward to measure the impact of each individual experiment on their treatment and control group, but how the experiments contribute to the average engagement across all users is unclear. In what follows we will assign costs Δ_l to individual experiments $l \in L$ relative to their impact by decomposing the overall loss in engagement due to the set of all active experiments L. Using concepts from cost sharing in game theory we can think about an experiment as a player. Several active players form a coalition, i.e., the specific treatment combination that a user gets exposed to.

We are interested in attributing the effect of an experiment to the overall population outcome, meaning that each player gets an assigned cost for being part of a coalition. We want to assess the loss compared with the baseline μ_0 that corresponds to the absence of experimentation (i.e., the control group). The overall loss is then denoted by $\Delta = \bar{Y} - \mu_0$ and we provide a decomposition such that $\Delta = \sum_{l \in L} \Delta_l$ where different experiments might interact with each other. This question is crucial if we want to check overall effects by experimentation, limit impact on the user experience and attribute the footprint fairly to stakeholders that run experiments. Our notion of fairness is derived from the concept of Shapley values, see [34].

Contributions. Our contributions are the following.

- We suggest two approaches based on causal inference and cost sharing games to attribute the impact of interacting treatments to average population level outcomes. We illustrate that a marginal perspective, looking at individual experiments only, would lead to misleading conclusions.

- We demonstrate the effectiveness of our approach in a large online user study on the Amazon Music platform as well as in a controlled, synthetic setting.
- We suggest an extension to conditional cost sharing that allows to predict shared impact in a counterfactual setting.

The rest of this paper is structured as follows. In Section 2 we review related work. Section 3 introduces background material and our suggested methodology. Section 4 details the experimental setup and highlights our results. Finally, Section 5 discusses our results and concludes.

Notation	Definition	Notation	Definition
Y_i	Outcome	$\bar{Y}$	Population average
$D_i(T)$	Treatment indicator	T_i	Treatment
μ_T	Average effect	Δ	Total cost
Δ_l	Attributed cost	$e(T_i, x_i)$	Propensity
X_i	Covariates	l	Active experiment

Table 1: Notation

2 RELATED WORK

Our work can be divided along three lines: (a) causal inference approaches for multivalued treatments; (b) measurement of the efficiency of ads on platform business, approaches to study parallel experiments and cost sharing games in game theory; and finally (c) trade-offs in recommender systems and multi-sided platforms.

Multiple simultaneous treatments in causal inference. Studying the effect of multivalued treatments goes back to at least the seminal work of [18] on the generalized propensity score. Estimating multivalued treatments effects [13, 17, 25, 36] has since then received substantial attention in the field of econometrics and biostatistics [26, 29]. Another perspective on analyzing simultaneous treatments is using experimental design approaches, see [11, 33]. In our work we make extensive use of this methodology to quantify the impact of experiments.

Measuring of ad effectiveness and cost sharing games. Due to its economic importance, the measurement of ad effectiveness has become a major field of application for causal inference, see for example recent work by [14, 22] and [24] for a perspective on parallel experimentation. We take a holistic perspective on the problem as we want to measure and disentangle combined impact of several experiments such as showing ads and promotions. We assess the impact of the experiments on the combined user experience which could be contrary to the aims of the experiment providers. Cost sharing games assign a value to contributing players and go back to the introduction of Shapley values by [40]. See also [4, 5, 20] for more background. Attributing treatment effects to various marketing channels in online advertising has been approached by combining cost sharing approaches with causal inference, see [41]. Our approach is distinct from the marketing attribution problem due to the parallel nature of the experiments.

Trade-offs in recommender systems. Optimizing engagement along other business objectives has become a crucial topic for multi stakeholder recommender systems, [1, 43]. Typical applications consist in automatic allocation of sponsored search [28, 48] or multi-relevance ranking [32] employing techniques from constraint optimization. Multi-objective optimization in market places [12, 23, 31] has also seen growing interest. For a recent line of work on the multi-sided and multi-objective nature of online platforms see, e.g., [8, 9, 30, 35]. In what follows we provide a perspective that splits impact fairly between different stakeholders and thus contributes to the understanding of platforms and their multi-sided nature.

3 BACKGROUND AND SUGGESTED APPROACH

Our aim is to decompose the average observed outcome $\bar{Y}$ among the different parties l that run experiments. We want to achieve this both in an empirical and counterfactual fashion. We consider the case of RCTs, where exposure to treatments is perfectly randomized, and observational studies, where treatments are potentially confounded. We employ the potential outcomes framework by [39]. For a recent survey see [47] or as major reference [19].

3.1 Causal Inference Techniques

We introduce required notation and the methods that we use to identify treatment effects.

Notation. We denote our outcome variable as $Y_i(T_i)$, where every unit of observation i can receive $2^{|L|}$ different treatments $T_i \in \mathcal{P}(L)$, which informs which treatment observation i receives, and $\mathcal{P}(L)$ is the power set denoting active experiments $l \in L$. Hence the experiments $l \in T$ inform us which experiment is part of a treatment. The empty set $\emptyset$ corresponds to the baseline μ_0, namely the control group. Pre-treatment covariables (i.e., measured before treatment is assigned) are denoted $X_i \in \mathbb{R}^d$. We define a treatment indicator $D_i(T)$ as

$$D_i(T) = \begin{cases} 1, & \text{if unit } i \text{ receives treatment } T \\ 0, & \text{otherwise} \end{cases}.$$

The observed outcome Y_i is then written as $Y_i = \sum_{T \in \mathcal{P}(L)} Y_{i,T} D_i(T)$, using the shorthand $Y_{i,T} = Y_i(T_i)$. We denote the expected outcome as $\mu_T = \mathbb{E}\left[Y_{i,T}\right]$, and we are interested in the population average treatment effect given as $\text{ATE}_{T,0} = \mu_T - \mu_0$, or more generally treatment comparisons of the form $\text{ATE}_{T,S} = \mu_T - \mu_S$. We define the lift over the baseline (in %) as $\text{lift}_{T,0} = \text{ATE}_{T,0}/\mu_0 \times 100$.

Assumptions. Following the definition in [47] the usual identification assumptions are

- Stable unit treatment value assumption (SUTVA): the potential outcome for a unit does not vary with treatments assigned to other units. There are no different versions of the treatment.
- Ignorability: Given the covariables X_i, the treatment assignment T_i is independent of the potential outcome, i.e., $T_i \perp Y_i(T_i)|X_i$
- Positivity: for any value of X, treatment is stochastic, i.e. $\mathbb{P}(T_i|X_i = x_i) > 0 \; \forall \; T_i, x_i$.

Propensity modeling. At the heart of most techniques in causal inference lies the propensity score [38] that is defined as the probability of receiving treatment:

$$e(T_i, x_i) = \mathbb{P}(T_i|X_i = x_i). \tag{1}$$

The propensity scores quantifies the fact that receiving treatment might depend on characteristics of the observation units. The propensity score can be modeled using, e.g., a multinomial logistic regression or non-parametric models. We denote the estimated propensity score by $\hat{e}(T_i, x_i)$.

Mean treatment. As a first approach for measuring treatment effects we introduce an estimator based on population averages. In the case of RCTs the average outcome can be estimated as

$$\hat{\mu}_T^{mean} = \sum_{i=1}^{n} D_i(T) Y_i / \sum_{i=1}^{n} D_i(T). \tag{2}$$

Inverse propensity weighting. As a second approach for estimating treatment effects under confounding, we use inverse propensity score weighting [16]. Under the ignorability of treatment assumption we estimate the expected outcome of treatment T by

$$\hat{\mu}_T^{IPS} = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i(T) Y_i}{\hat{e}(T_i, x_i)}. \tag{3}$$

Estimation of treatment effects. We estimate the average treatment effect over the baseline as $\hat{ATE}^{\bullet}_{T,0} = \hat{\mu}^{\bullet}_T - \hat{\mu}^{\bullet}_0$, where $\bullet$ denotes the methods {mean, IPS} as defined above. For comparing the effect of treatment T over S, we use $\hat{ATE}^{\bullet}_{T,S}$.

Marginal effects. Our exposition so far models the potential interaction of all experiments. We also estimate marginal effects, that ignore other running experiments. They are obtained using binary treatments of the form $T_{i,l} \in \{0, 1\}$.

3.2 Cost Sharing Games And Our Suggested Approach

The introduced approaches estimate the impact of different experiment combinations, but do not assign the individual contribution of l to the treatment T. The goal of our paper is to share the total impact over the baseline, i.e. $\mathbb{E}[\bar{Y}] - \mu_0$, among all contributors. This problem is known in the field of game theory as cost sharing game. Transcribed to our setting, the term *player* corresponds to active experiments. Players form *coalitions* (i.e., specific treatments). The term *grand coalition* denotes the set of all experiments. In a cooperative game with $|L|$ players each player $l \in L$ is assigned a value $\phi_l(v)$ for the game v. If players form a coalition $S \subset L$ this results in a cost of the coalition $v(S)$. Cost sharing mechanism should satisfy the following (axiomatic) properties for the assigned cost $\phi_l(v)$:

(a) Symmetry: it does not matter in which order the players l are numbered.
(b) Balanced budget: $\sum_{l \in L} \phi_l(v) = v(L)$ the sum of the individual values should equal the total outcome.
(c) Null player: $\phi_l(v) = 0$ if $l = \emptyset$, a player that does not contribute to the value of the game must have a null contribution.
(d) Additivity: for two games v, w $\phi_l(v) + \phi_l(w) = \phi_l(v + w)$.

The only cost sharing mechanism that satisfies all of the above criteria is the Shapley value [40]. See [34] for its connection with fairness and distributive justice. It is defined as

$$\phi_l^L(v) = \sum_{S \subset L - l} \frac{|S|!(|L| - |S| - 1)!}{|L|!} [v(S \cup l) - v(S)],$$

where we make explicit the dependence on the grand coalition L. One distinctive feature of Shapley values is the use of marginal contributions $v(S \cup l) - v(S)$. We now suggest two approaches to construct decompositions of the form $\mathbb{E}[\bar{Y}] - \mu_0 = \sum_{l \in L} \Delta_l$, that satisfy the balanced budget condition (b) as well as a perspective conditional on covariate values. The expectation of the average observed outcome is decomposed in a weighted sum of contributions: $\mathbb{E}[\bar{Y}] = \sum_{T \in \mathcal{P}(L)} \mu_T \times \mathbb{P}(T)$, where μ_T and $\mathbb{P}(T) := \mathbb{E}_X[e(T, X)]$ can be estimated using the techniques introduced before. The cost of a coalition is given as $v(S) = \mu_S - \mu_0$.

Weighted Shapley cost sharing. We suggest weighted Shapley cost sharing as solution to the cost sharing problem. Our solution is based on the concept of Shapley values that takes into account that not all experiments are active in parallel. The Shapley value conditional on the treatment T is $\phi_l^T(v)$, meaning that we only consider the subsets of coalitions up to T for computing this value. A weighted decomposition is then given as

$$\tilde{\Delta}_l = \sum_{T \in \mathcal{P}(L)} \mathbb{P}(T) \phi_l^T(v). \tag{4}$$

It is easily shown that weighted Shapley values are budget balanced. Note that if $l \notin T$, the null player property guarantees that $\phi_l^T(v) = 0$. Thus, this approach inherits all the favorable properties of Shapley values.

Weighted average cost sharing. As an alternative approach we suggest to divide the impact of treatment T equally among individual experiments l contributing to T, an approach corresponding to average cost sharing [46]. If a specific experiment l is part of a treatment, i.e. $l \in T$, we compute the impact of l on the total outcome via

$$\Delta_l = \sum_{T \in \mathcal{P}(L)} [\mu_T - \mu_0] \times \mathbb{P}(T) \frac{\mathbb{1}\{l \in T\}}{|T|}, \tag{5}$$

where $|T|$ is the number of active experiments inside the treatment T. The impact of l is thus the weighted impact over the baseline. The suggested decomposition has the balanced budget property but has the inconvenience that null players are not necessarily ignored. This makes this approach potentially unfair compared with weighted Shapley cost sharing. See also [5, 20] for further cost sharing and estimation approaches.

Conditional weighted cost sharing. As an extension to the two approaches we suggest a conditional perspective, where we look at conditional average treatment effects (CATE) [2] of the form $\mu_T(x) = \mathbb{E}[Y_i(T_i) | X = x]$. This idea provides insights on the impact for subgroups of the population. Estimating the CATE requires a model that predicts the outcome at the given covariate value $X = x$. In combination with the propensity score $e(T, x)$ we then define conditional cost sharing of the form $\Delta(x) = \sum_{l \in L} \Delta_l(x)$, which quantifies the impact on a specific subgroup of the population. The conditional weighted cost sharing approach can both be used in combination with Shapley or average cost sharing.

Approximating Shapley values. Our discussion so far assumed that the value of all possible coalitions are observed. In the case of missing combinations approximation techniques such as [27] could be used. The same idea can be used if the number of parallel

experiments gets large, since the exact computation of the Shapley value has a exponential runtime which can be prohibitive. We leave the investigation of this for future work.

4 EXPERIMENTS AND RESULTS

Synthetic experiment. We illustrate the difference between weighted Shapley cost sharing and average cost sharing in a synthetic experiment. See the Appendix for more details. In this experiment we deliberately introduce confounding by making treatment assignment dependent on the covariates X_i. Figure 1 illustrates the result: the approach based on IPS weighting, has less variance and yields more precise estimates compared to a naive strategy based on sample means. The approach based on average cost sharing results in all attributed costs watered down towards 0 and makes detecting significance more difficult. The weighted Shapley cost, however, allows a clear disentanglement when combined with IPS weighting.

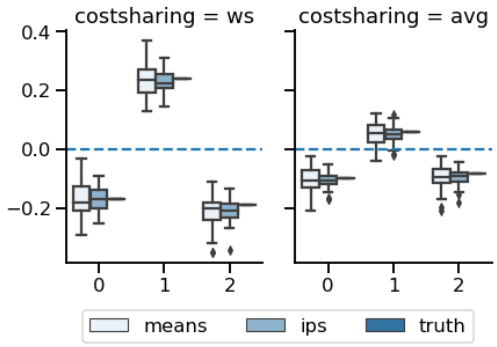

Figure 1: Shapley cost sharing (left) and average cost sharing (right) in a synthetic experiment. The shared cost is shown on the y-axis, the attributed experiment (numbered by 0,1,2) is indicated by the x-axis. The data has been simulated under covariate confounding which explains the higher error of the means estimator. The average cost sharing disentangles the contributions of the experiments less. The ground truth is indicated by a straight line next to the boxplots. Average cost sharing waters the contributions down to 0 and hence differences in the impact are not properly attributed.

Real world experiment. As a real world experiment we show results obtained on the Amazon Music platform using millions of observations. We studied three different experiments that were run in parallel for Amazon Music users in 7/2021 over two weeks. The experiments consisted in showing new editorial music content to a subset of users where some users saw the respective treatment and others the control. Treatments were perfectly randomized (checked using covariate balance). Running several experiments at once led to a reduction in −1.27% of overall CTR compared with no exposure to the three experiments. However, the contributions to the overall reduction was mostly due to a single experiment as Figure 2 illustrates. Experiment 0 lead to a loss on a small group, though this effect disappears when the experiment is active alongside the two other experiments (1,2). This interaction is not identified, when looking at marginal contributions only (see Table 2), where a negative impact is indicated, but not significant for any of the experiments. This is due to a dominance of the treatment where all experiments are active (around 66% of the population). Using an average cost sharing approach does not solve the problem, as all contributions are watered down to towards 0, without a significant indication of negative impact. The only approach that draws a clear picture is the one based on weighted Shapley values. Here, experiment 0 is significantly negative, whereas experiment 1 has a significantly positive lift on the overall experience (see Table 2). As experiment 0 is negatively contributing to the overall user experience, a resulting decision would be to disable it.

Table 2: Shared cost and marginal impact for the experiments on Amazon Music. Values measured as lift in %, the parentheses contain 95% confidence intervals. Bold values indicate significance.

	Exp. 0	Exp. 1	Exp. 2
Average cost	-0.59	-0.24	-0.44
(%)	(-1.35, 0.21)	(-0.97, 0.64)	(-1.26, 0.32)
Marginal Impact	-1.33	-0.87	-1.28
(%)	(-3.56, 1.20)	(-3.12, 1.85)	(-3.56, 1.31)
Shapley cost	**-10.19**	**7.93**	0.99
(%)	**(-12.31, -7.90)**	**(4.45, 10.97)**	(-1.41, 2.78)

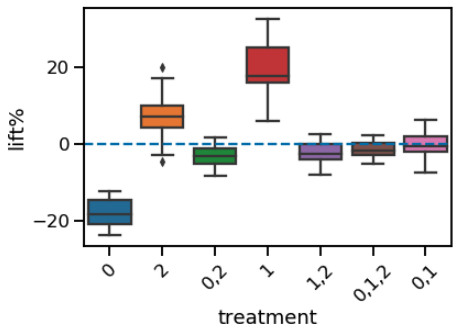

Figure 2: Impact of all experiment combinations (treatment, shown on the x-axis). We show the lift relative to the baseline in % (y-axis). The first experiment alone had a substantial negative impact, whereas the second experiment alone led to an uplift in engagement.

5 DISCUSSION AND CONCLUSION

We showed how to use causal inference techniques and cost sharing approaches to estimate and disentangle the effect of parallel experiments. Our weighted Shapley value approach attributes impact of parallel experiments and is a step necessary towards adaptive experimentation that limits impact beyond a given budget per experiment. The causal nature allows to predict impact on subgroups and hence creates a more fine grained perspective that goes beyond average impact. As future avenue we see the investigation of approximate Shapley values and missing data issues that arise if not all experiment combinations have materialized. We see benefit in investigating the combination with latest causal inference techniques such as doubly robust or double ML methods [6, 10].

ACKNOWLEDGMENTS

We would like to thank Moritz von Pein, Julian Dietz, Jan Malte Lichtenberg and Matej Jakimov for their support throughout the project.

REFERENCES

[1] Himan Abdollahpouri and Steve Essinger. [n. d.]. Multiple Stakeholders in Music Recommender Systems. ([n. d.]). arXiv:arXiv:1708.00120v1

[2] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. 2015. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics* 33, 4 (2015), 485–505.

[3] Joshua D. Angrist and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Number 8769 in Economics Books. Princeton University Press. https://ideas.repec.org/b/pup/pbooks/8769.html

[4] Eric Balkanski and Yaron Singer. 2015. Mechanisms for fair attribution. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. 529–546.

[5] Eric Balkanski, Umar Syed, and Sergei Vassilvitskii. 2017. Statistical cost sharing. *arXiv preprint arXiv:1703.03111* (2017).

[6] Heejung Bang and James M Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61, 4 (2005), 962–973.

[7] Matthew Blackwell and Nicole E Pashley. 2020. Noncompliance and instrumental variables for 2 factorial experiments. (2020).

[8] Robin Burke and Himan Abdollahpouri. 2017. Patterns of Multistakeholder Recommendation. (2017). arXiv:1707.09258 http://arxiv.org/abs/1707.09258

[9] Robin Burke, Himan Abdollahpouri, Bamshad Mobasher, and Trinadh Gupta. 2016. Towards multi-stakeholder utility evaluation of recommender systems. *CEUR Workshop Proceedings* 1618 (2016).

[10] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. Double/debiased/neyman machine learning of treatment effects. *American Economic Review* 107, 5 (2017), 261–65.

[11] David Roxbee Cox and Nancy Reid. 2000. *The theory of the design of experiments*. CRC Press.

[12] Weicong Ding, Dinesh Govindaraj, and S. V.N. Vishwanathan. 2019. Whole page optimization with global constraints. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2019), 3153–3161. https://doi.org/10.1145/3292500.3330675

[13] Ping Feng, Xiao Hua Zhou, Qing Ming Zou, Ming Yu Fan, and Xiao Song Li. 2012. Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine* 31, 7 (2012), 681–697. https://doi.org/10.1002/sim.4168

[14] Brett R. Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. 2019. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science* 38, 2 (2019), 193–205. https://doi.org/10.1287/mksc.2018.1135

[15] Miguel A Hernán and James M Robins. 2020. Causal inference: What If.

[16] Keisuke Hirano, Guido W Imbens, and Geert Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 4 (2003), 1161–1189.

[17] Liangyuan Hu, Chenyang Gu, Michael Lopez, Jiayi Ji, and Juan Wisnivesky. 2020. Estimation of causal effects of multiple treatments in observational studies with a binary outcome. *Statistical Methods in Medical Research* 29, 11 (2020), 3218–3234. https://doi.org/10.1177/0962280220921909

[18] Kosuke Imai and David A. Van Dyk. 2004. Causal inference with general treatment regimes: Generalizing the propensity score. *J. Amer. Statist. Assoc.* 99, 467 (2004), 854–866. https://doi.org/10.1198/016214504000001187

[19] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[20] Kamal Jain and Mohammad Mahdian. 2007. *Cost Sharing*. Cambridge University Press, 385–410. https://doi.org/10.1017/CBO9780511800481.017

[21] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery* 18, 1 (2009), 140–181.

[22] Akos Lada, Alexander Peysakhovich, Diego Aparicio, and Michael Bailey. 2019. Observational data for heterogeneous treatment effects with application to recommender systems. *ACM EC 2019 - Proceedings of the 2019 ACM Conference on Economics and Computation* (2019), 199–213. https://doi.org/10.1145/3328526.3329558

[23] Xiao Lin, Hongjie Chen, Changhua Pei, Fei Sun, Xuanji Xiao, Hanxiao Sun, Yongfeng Zhang, Wenwu Ou, and Peng Jiang. 2019. A pareto-eficient algorithm for multiple objective optimization in e-commerce recommendation. *RecSys 2019 - 13th ACM Conference on Recommender Systems* (2019), 20–28. https://doi.org/10.1145/3298689.3346998

[24] Xiliang Lin, Harikesh S. Nair, Navdeep S. Sahni, and Caio Waisman. 2019. Parallel experimentation in a competitive advertising marketplace. *arXiv* (2019), 1–51. arXiv:1903.11198

[25] Ariel Linden, S. Derya Uysal, Andrew Ryan, and John L. Adams. 2016. Estimating causal effects for multivalued treatments: A comparison of approaches. *Statistics in Medicine* 35, 4 (2016), 534–552. https://doi.org/10.1002/sim.6768

[26] Michael J. Lopez and Roee Gutman. 2017. Estimation of causal effects with multiple treatments: A review and new ideas. *Statist. Sci.* 32, 3 (2017), 432–454. https://doi.org/10.1214/17-STS612 arXiv:1701.05132

[27] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. 2013. Bounding the estimation error of sampling-based Shapley value approximation. *arXiv preprint arXiv:1306.4265* (2013).

[28] Edward C. Malthouse, Khadija Ali Vakeel, Yasaman Kamyab Hessary, Robin Burke, and Morana Fudurić. 2019. A multistakeholder recommender systems algorithm for allocating sponsored recommendations. *CEUR Workshop Proceedings* 2440 (2019).

[29] Daniel F. Mccaffrey, Beth Ann Griffin, Daniel Almirall, Mary Ellen Slaughter, Rajeev Ramchand, and Lane F. Burgette. 2013. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine* 32, 19 (2013), 3388–3414. https://doi.org/10.1002/sim.5753

[30] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace. (2018), 2243–2251. https://doi.org/10.1145/3269206.3272027

[31] Rishabh Mehrotra, Niannan Xue, and Mounia Lalmas. 2020. Bandit based Optimization of Multiple Objectives on a Music Streaming Platform. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3224–3233.

[32] Michinari Momma, Alireza Bagheri Garakani, and Yi Sun. 2019. Multi-objective relevance ranking. *CEUR Workshop Proceedings* 2410 (2019).

[33] Douglas C Montgomery. 2017. *Design and analysis of experiments*. John wiley & sons.

[34] Hervé Moulin. 2004. *Fair division and collective welfare*. MIT press.

[35] Phong Nguyen, John Dines, and Jan Krasnodebski. 2017. A Multi-Objective Learning to re-Rank Approach to Optimize Online Marketplaces for Multiple Stakeholders. (2017). arXiv:1708.00651 http://arxiv.org/abs/1708.00651

[36] Hui Nian, Chang Yu, Juan Ding, Huiyun Wu, William D Dupont, Tebeb Gebretsadik, Tina V Hartert, Pingsheng Wu, Hui Nian, Chang Yu, Juan Ding, Huiyun Wu, and William D Dupont. 2019. Performance evaluation of propensity score methods for estimating average treatment effects with multi-level treatments. 4763 (2019). https://doi.org/10.1080/02664763.2018.1523375

[37] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89, 427 (1994), 846–866.

[38] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[39] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66, 5 (1974), 688.

[40] Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.

[41] Raghav Singal, Omar Besbes, Antoine Desir, Vineet Goyal, and Garud Iyengar. 2019. Shapley meets uniform: An axiomatic framework for attribution in online advertising. In *The World Wide Web Conference*. 1713–1723.

[42] Dan Siroker and Pete Koomen. 2013. *A/B testing: The most powerful way to turn clicks into customers*. John Wiley & Sons.

[43] Andrew Stanton, Akhila Ananthram, Congzhe Su, and Liangjie Hong. 2019. Revenue, Relevance, Arbitrage and More: Joint Optimization Framework for Search Experiences in Two-Sided Marketplaces. (2019). arXiv:1905.06452 http://arxiv.org/abs/1905.06452

[44] Adith Swaminathan and Thorsten Joachims. 2015. The self-normalized estimator for counterfactual learning. In *advances in neural information processing systems*. Citeseer, 3231–3239.

[45] Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 17–26.

[46] Yair Tauman. 1988. The Aumann-Shapley prices: a survey. *The shapley value* (1988), 279.

[47] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2020. A survey on causal inference. *arXiv preprint arXiv:2002.02770* (2020).

[48] Xiangyu Zhao, Xudong Zheng, Xiwang Yang, Xiaobing Liu, and Jiliang Tang. 2020. Jointly Learning to Recommend and Advertise. July 2017 (2020). arXiv:2003.00097 http://arxiv.org/abs/2003.00097

A APPENDIX

In our appendix we provide more details on causal inference as well as our experimental set-up.

A.1 More details on causal inference

For the sake of completeness we introduce more involved causal inference methods that can be used in combination with our cost sharing approach.

Regression adjustment. Assuming a linear relationship between the outcome variable and pre-treatment covariates we can use a linear model to directly estimate the treatment effect. This approach, commonly used in econometrics [3], has the advantage of directly providing estimates for the uncertainty of the treatment coefficients. A clear downside of this approach, however, is that the underlying assumption can be overly restrictive and a violation of the presupposed linearity can result in biased inference. The linear regression adjustment model is defined as

$$Y_i = \sum_T \alpha_T D_i(T) + \beta_0 + \sum_{j=1}^{J} \beta_j X_{i,j} + \epsilon_i, \tag{6}$$

where ϵ_i is an error term, $X_{i,j}$ are our pre-treatment covariables. Counterfactual prediction for individual values of K can be obtained using Equation (6) for prediction of the form $\hat{m}(T, x_i) \approx \mathbb{E}[Y_i(T)|X_i = x_i]$, which can be used for the computation of the conditional weighted Shapley values. An estimate for the treatment effect is obtained as $\hat{\mu}_T^{RA} = 1/n \sum_{i=1}^{n} \hat{m}(T, x_i)$.

Self normalized IPS. The self normalized IPS estimator makes the IPS method more robust by reducing the variance of small weights [19, 44]. The corresponding estimator is biased but consistent. Its form is given as

$$\hat{\mu}_T^{snIPS} = \hat{\mu}_T^{IPS} / \sum_{i=1}^{n} \frac{1}{n} \frac{D_i(T))}{\hat{e}(T_i, x_i)}. \tag{7}$$

Doubly robust estimator. The class of doubly robust estimators [37] combines the propensity score estimation with the regression adjustment. This makes this class of estimators correct if either the propensity score adjustment or the regression adjustment is incorrect (but not both). A widely used version of this estimator is given as

$$\hat{\mu}_T^{DR} = \sum_{i=1}^{n} \left(\hat{m}(T, x_i) + \frac{D_i(T)(Y_i - \hat{m}(T, x_i))}{\hat{e}(T_i, x_i)} \right) / \sum_{i=1}^{n} D_i(T). \tag{8}$$

Here $\hat{m}(T, x_i)$ are the predictions of the regression adjustment model of Equation (6) for the treatment set to T for covariate x_i.

The doubly robust estimator can be estimated using a two step approach where first we estimate the multivariate propensity model by regressing the treatment on the pretreatment covariates. Then as a second step we regress the observed outcomes for treatment T on the pretreatment covariates. The predicted outcome for the regression adjustment and the propensity score model are then used to compute the estimator $\hat{\mu}_T^{DR}$. This estimator has typically a higher variance than the RA estimator, if its underlying model is correct, but in practice the doubly robust property is often worth this loss.

A.2 Details on the experiments

Standard error estimation. Confidence intervals and standard errors are computed using bootstrapping where we resample datasets 200 times with replacement (see, for example, [15]). Then, we compute the estimated lift using the different methods introduced in Section 3.1.

Multivariate estimation of treatment effects and lift. We estimate treatments effects and lift using a multivariate approach. Here, we estimate jointly the effect of experiment as well as their interactions. The number of treatments is $2^L - 1$, where L denotes the number of experiments. The empty set $\emptyset$ serves as baseline. The propensity score is derived from a multinomial model with $2^{|L|}$ different classes.

Univariate estimation of treatment effects and lift. We estimate treatments effects and the associated lift using an univariate approach looking at each experiment l individually. We focus on the impact on the experiment level without taking into account potential interactions. We therefore term this approach *marginal* effect estimation, as we estimate the causal effect of a binary treatment at the experiment level.

Synthetic experiment. The underlying data is generated by fixing first two parameter β_1 and β_2, then we simulate a normal distributed covariate vector X. Then X and β_1 generate treatment assignment in a multinomial model resulting in T. X and β_2 are then used to add more confounding. We provide python code for the data generation below.

```
import numpy as np
import pandas as pd
np.random.seed(42)

class GenerateSyntheticSample(object):
    """
    function that generates synthetic sample
    """
   def __init__(self, dim, m_treatments=3, rct=False, seed=None):
        self.dim = dim
        self.m_treatments = m_treatments
        self.rct = rct
       # generate random treatment effects between -1,1
        np.random.seed(42)
        self.tau = 2*np.random.uniform(
          size=(2**self.m_treatments-1))-1
        np.random.seed(None)
        self.seed = seed

        beta_1 = np.linspace(-dim, dim, num=dim)
      self.beta_1 = beta_1 / np.sum(beta_1 ** 2) ** 0.5
        beta_2 = np.linspace(dim, -dim, num=dim)
      self.beta_2 = beta_2 / np.sum(beta_2 ** 2) ** 0.5

    def generate_sample(self, n):
        if self.seed:
            np.random.seed(self.seed)

        x = np.random.normal(size=(n, self.dim))
```

```
if self.rct:
    p = np.ones((n, self.m_treatments)) * 0.5
else:
    p = np.zeros((n, self.m_treatments))
    for mi in range(self.m_treatments):
        p[:, mi] = 1.0 /
   (1.0 + np.exp(-x.dot(self.beta_1)*(-1**mi)))

d = np.random.binomial(1, p)
treatment_strings = list(map(''.join,
    d.astype(int).astype(str)))
d_all = pd.get_dummies(treatment_strings,
drop_first=True)

self.feature_names = d_all.columns
        y = x.dot(self.beta_2) +
         d_all.dot(self.tau) + np.random.normal(size=n)
        weights = np.ones(n)
        return y, x, d, weights

syntheticsamplegenerator =
  GenerateSyntheticSample(5, rct=False)
y, x, d, weights =
  syntheticsamplegenerator.generate_sample(10000)
```

Real world experiment. Our real world experiment uses data from Amazon Music. In order to not disclose sensitive business information, we refrain from giving exact details on the experiment.