Supplemental Material for "Statistical Challenges in Online Controlled Experiments: A Review of A/B Testing Methodology"

SM1 Stratified Sampling and CUPED

Assume there exist K strata dividing the population Ω, where every stratum has mean and variance $(\mu_k,\ \sigma_k^2)$, and each unit i falls into the k^{th} strata with unknown probability w_k such that $\sum_{k=1}^{K} w_k = 1$. With data obtained via stratified sampling, it is well-known that one may construct an unbiased, weighted estimator of τ that has smaller variance than the standard difference-of-means estimator, presuming one has correctly estimated w_k and identified stratum that are correlated with Y (Acharya et al., 2013). As noted in Deng et al. (2013) and Xie and Aurisset (2016), many organizations have access to large amounts of data, which can simplify the process of identifying meaningful strata. However, estimating w_k is not straightforward, and the real-time nature of online experiments as well as the physical infrastructure of experimentation platforms also hinder accurate implementation of stratified random sampling. The primary challenge is to maintain equal representation of the strata while users are randomized to treatment and control. Xie and Aurisset (2016) propose a novel stratified sampling technique that involves defining one queue q for each strata k. Each q consists of multiple segments of fixed length. Depending on their strata, users are first assigned to a slot within a segment, then treatments are randomized within each segment.

Consequently, balanced allocation is only guaranteed within a segment. Moreover, if multiple machines each have their own q for strata k, as is the case in many large experimentation platforms, balanced randomization is even more difficult to achieve. Deng et al. (2013) show that CUPED is equivalent to stratified random sampling when the control variate is categorical, and is considered a post-experiment workaround for the practical difficulties of implementing stratified sampling in real-time. Xie and Aurisset (2016) compare their stratified sampling technique to CUPED, finding that CUPED prevails in terms of variance reduction. Practitioners continue to be interested in methods for stratified sampling with the aim of variance reduction, as well as identifying such strata in order to detect bugs or potential areas for targeted optimization.

SM2 Surrogate Outcomes for Long-term Treatment Effect Estimation

This literature generally begins with the following. Assume a potential outcomes setup with two samples, n_E (experimental) and n_O (observational), with binary indicator $G_i \in \{E, O\}$. The tuple $(W_i,\ S_i,\ X_i)$ is observed in the experimental group and $(Y_i,\ S_i,\ X_i)$ in the observational group, where S_i is an intermediate short-term outcome and X_i is a pre-treatment covariate (W_i may also be included in the observational group, see Athey et al. (2020a) and Imbens et al. (2022)). The goal is to estimate the average treatment effect of W_i on Y_i, which is nontrivial since Y_i is not observed in the experimental sample. The origins of this framework can be traced back to statistical literature regarding *surrogate outcomes*, used largely in biostatistics and econometrics (Prentice, 1989; Begg and Leung, 2000; Frangakis and Rubin, 2002; Ensor et al., 2016). The work by Athey et al. (2019) is one of the first papers that uses this framework for long-term effect estimation cited within the OCE community. The authors derive estimators of τ using S_i as driver metrics and assume W_i is not observable in O. They employ the "surrogate criterion", which requires that Y_i

be independent of W_i given the short-term outcomes. It is straightforward to see that the approach in Hohnhold et al. (2015) is a special case of this approach, where S_i is comprised of the learned click-through-rates and short-term revenue, Y_i is long-term revenue, and the necessary conditions for estimating τ are unverified but implicitly assumed.

In practice, the surrogate criterion is notoriously tricky to satisfy. Athey et al. (2020b) relax this assumption by only requiring that Y_i is independent of W_i conditional on a *set* of surrogates, rather than on each individual surrogate. In perhaps one of the earliest publications using statistical surrogacy to estimate long-term effects specifically in OCEs, Cheng et al. (2020) show that one can relax the surrogacy assumption by extending this framework to incorporate sequential testing. There is also evidence that some tech companies such as Facebook have used statistical surrogacy (Gupta et al., 2019), although it appears that too many surrogates may severely hamper interpretability. Recent work has shifted away from the surrogate criterion. Athey et al. (2020a) let W_i be seen in the observational sample and estimate the treatment effect on S_i in both samples, using the difference to adjust the ATE estimates. Imbens et al. (2022) consider a similar context and demonstrate how to account for unmeasured confounding variables that impact treatment, short-term, and long-term outcomes. Van Goffrier et al. (2023) point out that surrogate methods that assume there are no unobserved confounders in the observational data may not be a practically useful. As an alternative they propose an instrumental variables approach to estimate a long-term effect by combining regression residuals with short-term experimental outcomes. Further exploration of combining short-term experimental data with observational data to estimate long-term effects may show promise with respect to OCE applications.

SM3 Network A/B Tests

OCEs where the experimental units are subject to network exposure are known as *network A/B tests*, where users and the connections among them are modeled by a network $\mathcal{G}$, with

$n \times n$ adjacency matrix $\mathbf{A} = [A_{ij}]$. In most OCE settings, $\mathbf{A}$ is assumed to be fixed and observable, although situations where this is not the case are also considered (Egami, 2017). The goal of estimating the ATE remains of primary interest. However, when SUTVA is violated, standard randomization schemes and estimators tend to ignore the network effect, which typically produces biased estimates of τ. Consider the following example: suppose the response $Y_i = \alpha + \beta W_i + \gamma S_i + \varepsilon_i$ is linearly related to the treatment effect β and network spillover effect γ, where S_i is the proportion of i's neighbors that received treatment. The ATE is therefore $\beta + \gamma$, since $\mathbb{E}[S_i|W_i = 1] = 1$ and $\mathbb{E}[S_i|W_i = 0] = 0$. Under the usual balanced randomization, however, $\mathbb{E}[S_i] = 0.5$ for both treatment and control groups, thus the expected value of the usual difference of means estimator $\hat{\tau}$ is β, which has a bias of γ. Generally, the exact form of the ATE depends on the assumed structure of $\mathcal{G}$ and definition of S_i; similarly for the form and bias of $\hat{\tau}$. Thus, there are two major problems in network A/B testing that current research aims to address: (1) modeling and estimating the network spillover effect, and (2) optimal treatment allocation for producing unbiased estimates of τ in the presence of network interference. Reviewing work in these areas is the focus of the following subsection

A commonly proposed approach for dealing with network effects in OCEs is to randomize treatments with *graph cluster randomization* (Karrer et al., 2021; Eckles et al., 2014; Gui et al., 2015; Saveski et al., 2017; Sangho Yoon, 2018; Zhou et al., 2020; Ugander et al., 2013). With cluster-based randomization, the network is partitioned into subgroups or *clusters*, such that edge connectivity within clusters is higher than between clusters. Network partitioning, also known as community detection in network science, is a well-researched area, with most OCEs using established graph clustering algorithms as found in Newman (2006), Leskovec et al. (2010), and Mucha et al. (2010) and Stanley et al. (2016). Treatments are then randomized to users at the cluster level with the standard difference of means estimator, a common choice for estimating the ATE. Eckles et al. (2014) explore several linear models for relating user response to the network effect, and perform a suite of simulations that show graph cluster

randomization reduces bias when compared to naive random allocation. They also provide a theorem that shows the bias from network effects will always be less than or equal to the bias from random allocation, assuming $Y_i = \alpha + \beta W_i + \gamma S_i + \varepsilon_i$. Gui et al. (2015) draw from this work, modeling the response as $Y_i = \alpha + \beta W_i + \gamma \sum_{j=1}^{n} A_{ij} W_j + \eta \sum_{j=1}^{n} A_{ij} Y_j / d_i$, where d_i is the degree of node i, γ is the spillover effect, and η describes how users tend to exhibit behavior similar to their neighbors'. They showed that with a network sampled such that clusters are "balanced", where clusters are all equal in size, one can eliminate the bias in $\hat{\tau}$. Their new algorithm for balanced cluster-based randomization was empirically shown to reduce bias, although theoretical justification was not provided. To address the question of how to detect when the spillover effect is present, Saveski et al. (2017) develop a model-free two stage cluster-randomization design for testing for the presence of SUTVA violations, and Athey et al. (2018) derive exact p-values for nonsharp null hypotheses of no spillover effects. Recent work by Karrer et al. (2021) utilizes imbalanced clusters with a regression-adjusted estimator, along with a post-analysis framework that is also used to detect network effects.

While Gui et al. (2015) use a common framework for OCE applications, the linear model assumption is known to be quite restrictive, particularly for network applications. Basse and Airoldi (2018) specifically study the drawbacks of traditional parametric assumptions for modeling network effects. Some practioners instead use *network exposure models* to model the spillover effect (Backstrom and Kleinberg, 2011; Katzir et al., 2012). Network exposure models define a set of conditions for each i under which the spillover effects from i's neighbors are the same. For example, the *neighborhood exposure model* from Backstrom and Kleinberg (2011) and Gui et al. (2015) estimates τ with $\frac{1}{|N_1^\theta|} \sum_{i \in N_1^\theta} Y_i - \frac{1}{|N_0^\theta|} \sum_{i \in N_0^\theta} Y_i$, where σ_i is the percent of neighbors of i that received treatment, $N_1 = \{i : W_i = 1, \sigma_i \geq \theta\}$, $N_0 = \{i : W_i = 0, \sigma_i \leq 1 - \theta\}$, and $\theta \in [0, 1]$. With network exposure models, one need not make explicit assumptions about how the spillover effect relates to the response, although the corresponding ATE estimators tend to be more complex. Ugander et al. (2013) catalogue the various network exposure models that have been commonly adopted in the literature

(Eckles et al., 2014; Gui et al., 2015; Saveski et al., 2017).

While cluster-based randomization approaches are commonly used in practice, the limitations of this method are significant enough that researchers remain interested in alternative approaches. First, because this approach uses clusters as the experimental units and cluster counts typically are far smaller than the total number of users, experiments under this approach tend to lack adequate power. To mitigate this, Saint-Jacques et al. (2019) propose sampling many "ego-networks", which are *small* clusters comprised of a central user and a carefully selected subset of their immediate neighbors. Second, the majority of online social networks are highly dense, making it extremely difficult to obtain reasonably isolated clusters that are representative of the true network. Nandy et al. (2020) avoid explicit model assumptions by defining $\mathcal{G}$ as a directed network of producers j and consumers i. Treatment intervention (r) is represented by rewiring edge probabilities by replacing the original p_{ij}^{base} with $p_{ij}^{(r)}$, where $p_{ij} = Pr(A_{ij} = 1)$. Nandy et al. (2020) use this setup to frame treatment allocation as an optimization problem, where treatments are randomized such that the effect from network exposure under the new treatment is as small as possible. Their method showed an improvement over cluster-based randomization in terms of bias of the ATE, particularly for highly dense networks. Note Nandy et al. (2020) and Saint-Jacques et al. (2019) and Gui et al. (2015) all assume that the network is known, where in fact it is highly possible there are unobserved covariates or network effects influencing network structure and user response. Bajari et al. (2021) employ the producer-consumer marketplace set-up to address interference without a network model. Rather, users are assumed to belong to a number of different populations that serve as indices for the outcomes and treatment assignments. Bajari et al. (2021) define a new class of experimental designs, *Multiple Randomization Designs*, that model the response as a tuple with elements corresponding to each population and randomize treatments at the tuple-level.

Despite the drawbacks of defining a parametric model for Y_i, there are inherit advantages to this approach, such as analyzing heterogeneity in the form of interactions or applying

conventional tools like censoring and stratification (Walker and Muchnik, 2014). Under this framework, a natural solution to the question of treatment allocation is optimal design of experiments theory. Optimal design refers to the general practice of choosing a design matrix from the space of potential candidates, $X \in \mathcal{X}$, according to various optimality criterion. In Parker et al. (2017), the response is modelled as $Y_i = \alpha + \tau_{t(i)} + \sum_{j=1}^{n} A_{ij}\gamma_{t(j)} + \varepsilon_i$, where $\tau_{t(i)}$ represents the treatment applied to i, assuming $k \in \{1, ...K\}$ treatments. A blocking parameter b_i can also be introduced to this model (Koutra, 2017). With this framework, Parker et al. (2017) and Koutra (2017) provide some interesting insights into what optimal designs for network A/B testing might look like, namely that unbalanced designs tend to be better at reducing the variance of $\hat{\tau}_j$ than balanced allocation. However, these models are rather unrealistic. Because they do not scale the spillover effect by the degree of node i, as the number of neighbors of i grows, $\sum_{j=1}^{n} A_{ij}\gamma_{t(j)} \to \infty$ as well, meaning the spillover effect completely dominates $\tau_{t(i)}$ for the large networks typically observed in OCEs. Parker et al. (2017) and Koutra (2017) also do not optimize for the ATE, instead considering optimal designs for only τ_j by minimizing the average variance of all pairwise treatment effects. Additionally, these optimal designs are chosen with an exhaustive search algorithm, which searches the entire space of $\mathcal{X}$, or K^n potential designs, before selecting X. Indeed, some of the designs obtained via search algorithm in Parker et al. (2017) were outperformed by randomly generating X. Pokhiko et al. (2019) and Zhang and Kang (2020) alternatively choose conditional auto-regressive models to mimic the network effect by correlating the response error of i with that of its neighbors. A strong limitation of this approach is this correlation is assumed to be the same across all nodes. Zhang and Kang (2020) address this issue by using Bayesian priors via simulation, but do not leverage network information in defining them.

SM4 Beyond This Review

We have presented literature that generally assumes a single treatment and control under a frequentist framework. While this setting describes an appreciable majority of OCEs, there is also growing interest in methodologies that extend beyond the scope of this review. Researchers aiming to circumvent limitations of the frequentist p-value have turned to Bayesian methods (Stucchio, 2015; Letham et al., 2019; Deng et al., 2016; Deng et al., 2021; Kamalbasha and Eugster, 2021; Hoffmann and Wagenmakers, 2021), including implementations of Bayes factor hypothesis testing (Deng, 2015) and tests for practical significance (Stevens and Hagar, 2022). Many practitioners have noted that the ATE itself is not a quantity of interest in several applications, e.g., when optimizing tail performance, and have begun to develop approaches using *quantile metrics* (Liu et al., 2019; Howard and Ramdas, 2019; Lux, 2018). *Multi-armed bandits* have been used to handle multiple treatments in online settings, with a focus on sequential decision-making and exposing more users to successful variants to increase reward (Liu et al., 2014; Issa Mattos et al., 2019; Birkett, 2019; Amadio, 2020; Lomas et al., 2016). Thompson sampling (Scott, 2010; Scott, 2015; Dimakopoulou et al., 2021) as well as contextual bandits (Li et al., 2010; Agarwal et al., 2016) have all been used in industry. Novel experimental designs have also been developed for purposes of increasing sensitivity in low-power settings; several of these were discussed in Section 6 in the context of mitigating interference. Another commonly used design, particularly in the context of experiments on search ranking algorithms, is *interleaving* (Radlinski and Craswell, 2013; Parks et al., 2017; Zhang et al., 2022). Rather than displaying results to a user from either a treatment algorithm or a control algorithm, this design involves interleaving search results from both the treatment and control algorithms. Thus, each user experiences both the treatment and control simultaneously, thereby providing additional information, yielding insights faster.

Although OCEs with multiple variants are reasonably common, full- and fractional-factorial experiments that emphasize estimation of main and interaction effects are uncom-

mon; Kohavi et al. (2009) and Georgiev (2019) argue that the added practical complexity of such experiments hurts development agility and is not worth the additional effort when interactions in practice are rare. They suggest that it is preferable to run multiple single-factor experiments concurrently, , and validate that there are no significant interactions between all pairs of experiments (Gupta et al., 2019). *Multivariate tests* (where the multiple variants are defined by the factorial enumeration of multiple factors' levels) *do* exist in this space (McFarland, 2012; Wildman, 2019), but the goal of the analysis is primarily to identify the optimal variant, *not* to estimate individual effects. Though multivariate tests are not as common as A/B or A/B/n tests, research in this area carries on (Sadeghi et al., 2019), with recent research in optimal design (Bhat et al., 2020; Basse et al., 2023; Bojinov et al., 2022), nonparametric estimators for panel experiments (Bojinov et al., 2021), and factorial designs for sequential testing (Haizler and Steinberg, 2020). How to avoid, identify, and estimate interactions between multiple concurrent experiments is also of great interest (Kohavi et al., 2009; Gupta et al., 2019; Chan, 2021).

Another important facet of OCEs outside the scope of this review is the issue of ethics (Gupta et al., 2019; Kohavi et al., 2020). As noted, the experimental units in OCEs are often people – human subjects – and so a salient concern is whether experiments involving them are ethical. Many OCEs test harmless interface changes, but there exist A/B tests that through *code* induce *deception*, thus named C/D tests (Benbunan-Fich, 2017; Kontotasiou, 2021). One example is Facebook's infamous *emotional contagion* experiment in which the sentiment of content shown in nearly 700,000 users' News Feeds was altered to determine whether this impacted their own emotions (Kramer et al., 2014). Another example is OKCupid's *power of suggestion* experiment in which matched users were told their compatibility was higher than what the matching algorithm predicted in order to investigate the impact of simply telling couples they're a good match (Rudder, 2014).

More recently, there were ethical questions (Singer, 2022) about a retrospective analysis of experiments run by LinkedIn from 2015 to 2019 (Rajkumar et al., 2022) in order to

understand the *strength of weak ties* social theory (Granovetter, 1973). These experiments engaged 20 million users and tested changes designed to improve the algorithm underlying the "People You May Know" (PYMK) feature. It is important to note that LinkedIn did not intentionally vary the proportion of weak and strong contacts suggested by PYMK (Belanger, 2022) but that these variations were side effects of experiments optimizing for other criteria. It is unknown if these changes have negatively or positively impacted users looking for job opportunities. Another context in which OCEs may have unintended side effects is digital labor platforms in today's "gig" economy. In this setting, the experimental units are typically the workers using the platform and researchers have found that continuous and concealed experimentation diminishes worker autonomy and satisfaction (Rahman et al., 2023).

The primary concern in these settings is informed consent; users generally do not know when they're being experimented on, nor do they necessarily have a way to opt out of such an experiment. They implicitly consent to such experimentation when they agree to a service's terms and conditions, however, whether such consent is *informed* is debatable (Benbunan-Fich, 2017). Academics involved in human subjects research will be familiar with institutional review boards (IRBs) and ethics clearance. Such formal oversight is generally absent in the private sector. However, Kohavi et al. (2020) do advocate for the establishment of processes that fulfill this purpose so that an experiment's risks and benefits are carefully considered, and transparent protocols for informed consent and drop-out are instated. The authors also advocate for tools, infrastructure, and processes to ensure data security and data privacy, another issue especially relevant in this day and age. See Kohavi et al. (2020) and Bojinov and Gupta (2022) for expanded discussions of identified data, anonymous data, re-identification, and differential privacy in the context of OCEs.

In contexts where a controlled experiment is unethical or infeasible, companies have turned to observational causal inference methods. For instance, Mozilla is interested in the impact of ad blocker installation on browser engagement (Miroglio et al., 2018); Netflix wants to quantify the cumulative effect of in-device promotions and out-of-home marketing

for a particular title (McFarland et al., 2018); Uber Eats is interested in understanding how delivery delays influence a user's future engagement with the platform, and Uber is interested in how ride bookings are impacted by surge pricing rates (Harinen and Li, 2019). In these cases, and others like them, a traditional OCE is not appropriate or not possible, so companies estimate causal impacts using methods like matching, regression discontinuity, interrupted time series, instrumental variables, and difference in differences, among others. Like OCEs, this is a rapidly growing area that merits a literature review of its own.

References

Acharya, Prakash, Saxena, and Nigam (2013). "Sampling: Why and how of it". *Indian Journal of Medical Specialties* 4.2, pp. 330–333.

Agarwal, Bird, Cozowicz, Hoang, Langford, Lee, Li, Melamed, Oshri, Ribas, et al. (2016). "Making contextual decisions with low technical debt". *arXiv preprint arXiv:1606.03966*.

Amadio (2020). *Multi-Armed Bandits and the Stitch Fix Experimentation Platform*. https://multithreaded.stitchfix.com/blog/2020/08/05/bandits/. (Accessed on 03/21/2022).

Athey, Chetty, and Imbens (2020a). *Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes*. DOI: 10.48550/arxiv.2006.09676.

Athey, Chetty, Imbens, and Kang (2020b). *Estimating Treatment Effects using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index*. arXiv: 1603.09326 [stat.ME].

Athey, Chetty, Imbens, and Kang (2019). *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely*. Tech. rep. National Bureau of Economic Research.

Athey, Eckles, and Imbens (2018). "Exact p-values for network interference". *Journal of the American Statistical Association* 113.521, pp. 230–240.

Backstrom and Kleinberg (2011). "Network bucket testing". *Proceedings of the 20th international conference on World wide web*. WWW '11. New York, NY, USA: Association for Computing Machinery, pp. 615–624. ISBN: 978-1-4503-0632-4. DOI: 10.1145/1963405.1963492.

Bajari, Burdick, Imbens, Masoero, McQueen, Richardson, and Rosen (2021). "Multiple Randomization Designs". *arXiv preprint arXiv:2112.13495*.

Basse, Ding, and Toulis (2023). "Minimax designs for causal effects in temporal experiments with treatment habituation". *Biometrika* 110.1, pp. 155–168.

Basse and Airoldi (2018). "Limitations of Design-based Causal Inference and A/B Testing under Arbitrary and Network Interference". *Sociological Methodology* 48.1. Publisher: SAGE Publications Inc, pp. 136–151. ISSN: 0081-1750. DOI: 10.1177/0081175018782569.

Begg and Leung (2000). "On the use of surrogate end points in randomized trials". *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 163.1, pp. 15–28.

Belanger (2022). *xperts debate the ethics of LinkedIn's algorithm experiments on 20M users*. https://arstechnica.com/tech-policy/2022/09/experts-debate-the-ethics-of-linkedins-algorithm-experiments-on-20m-users/.

Benbunan-Fich (2017). "The ethics of online research with unsuspecting users: From A/B testing to C/D experimentation". *Research Ethics* 13.3-4, pp. 200–218.

Bhat, Farias, Moallemi, and Sinha (2020). "Near-Optimal A-B Testing". *Management Science*. Publisher: INFORMS. ISSN: 0025-1909. DOI: 10.1287/mnsc.2019.3424.

Birkett (2019). *When to Run Bandit Tests Instead of A/B/n Tests*. CXL. Library Catalog: cxl.com. URL: https://cxl.com/blog/bandit-tests/ (visited on 06/16/2020).

Bojinov and Gupta (2022). "Online experimentation: Benefits, operational and methodological challenges, and scaling guide". *Harvard Data Science Review* 4.3.

Bojinov, Rambachan, and Shephard (2021). "Panel experiments and dynamic causal effects: A finite population perspective". *Quantitative Economics* 12.4, pp. 1171–1196.

Bojinov, Simchi-Levi, and Zhao (2022). "Design and analysis of switchback experiments". *Management Science*.

Chan (2021). *Embrace Overlapping A/B Tests and Avoid the Dangers of Isolating Experiments*. https://blog.statsig.com/embracing-overlapping-a-b-tests-and-the-danger-of-isolating-experiments-cb0a69e09d3. (Accessed on 03/21/2022).

Cheng, Guo, and Liu (2020). "Long-Term Effect Estimation with Surrogate Representation". *arXiv preprint arXiv:2008.08236*.

Deng (2015). "Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments". *Proceedings of the 24th International Conference on World Wide Web*. WWW '15 Companion. Florence, Italy: Association for Computing Machinery, pp. 923–928. ISBN: 978-1-4503-3473-0. DOI: 10.1145/2740908.2742563.

Deng, Li, Lu, and Ramamurthy (2021). "On Post-selection Inference in A/B Testing". *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2743–2752.

Deng, Lu, and Chen (2016). "Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing". *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pp. 243–252.

Deng, Xu, Kohavi, and Walker (2013). "Improving the sensitivity of online controlled experiments by utilizing pre-experiment data". *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*. the sixth ACM international conference. Rome, Italy: ACM Press, p. 123. ISBN: 978-1-4503-1869-3. DOI: 10.1145/2433396.2433413.

Dimakopoulou, Ren, and Zhou (2021). "Online Multi-Armed Bandits with Adaptive Inference". *Advances in Neural Information Processing Systems* 34.

Eckles, Karrer, and Ugander (2014). "Design and analysis of experiments in networks: Reducing bias from interference". *arXiv:1404.7530 [physics, stat]*. arXiv: 1404.7530.

Egami (2017). "Unbiased estimation and sensitivity analysis for network-specific spillover effects: Application to an online network experiment". *arXiv preprint arXiv:1708.08171*.

Ensor, Lee, Sudlow, and Weir (2016). "Statistical approaches for evaluating surrogate outcomes in clinical trials: a systematic review". *Journal of biopharmaceutical statistics* 26.5, pp. 859–879.

Frangakis and Rubin (2002). "Principal stratification in causal inference". *Biometrics* 58.1, pp. 21–29.

Georgiev (2019). *Statistical methods in online A/B testing*. Self-Published.

Granovetter (1973). "The strength of weak ties". *American journal of sociology* 78.6, pp. 1360–1380.

Gui, Xu, Bhasin, and Han (2015). "Network A/B Testing: From Sampling to Estimation". *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. Florence, Italy: International World Wide Web Conferences Steering Committee, pp. 399–409. ISBN: 978-1-4503-3469-3. DOI: 10.1145/2736277.2741081.

Gupta, Kohavi, Tang, Xu, Andersen, Bakshy, Cardin, Chandran, Chen, Coey, Curtis, Deng, Duan, Forbes, Frasca, Guy, Imbens, Saint Jacques, Kantawala, Katsev, Katzwer, Konutgan, Kunakova, Lee, Lee, Liu, McQueen, Najmi, Smith, Trehan, Vermeer, Walker, Wong, and Yashkov (2019). "Top Challenges from the First Practical Online Controlled Experiments Summit". *SIGKDD Explor. Newsl.* 21.1, pp. 20–35. ISSN: 1931-0145. DOI: 10.1145/3331651.3331655.

Haizler and Steinberg (2020). "Factorial Designs for Online Experiments". *Technometrics*, pp. 1–12. ISSN: 0040-1706, 1537-2723. DOI: 10.1080/00401706.2019.1701556.

Harinen and Li (2019). *Using causal inference to improve the uber user experience*. https://www.uber.com/en-CA/blog/causal-inference-at-uber/.

Hoffmann and Wagenmakers (2021). "Bayesian inference for the A/B test: Example applications with r and jasp". *PsyArXiv. June* 10.

Hohnhold, O'Brien, and Tang (2015). "Focusing on the Long-term: It's Good for Users and Business". *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. Sydney, NSW, Australia: Association for Computing Machinery, pp. 1849–1858. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2788583.

Howard and Ramdas (2019). "Sequential estimation of quantiles with applications to A/B-testing and best-arm identification". *arXiv preprint arXiv:1906.09712*.

Imbens, Kallus, Mao, and Wang (2022). *Long-term Causal Inference Under Persistent Confounding via Data Combination*. DOI: 10.48550/ARXIV.2202.07234.

Issa Mattos, Bosch, and Olsson (2019). "Multi-armed bandits in the wild: Pitfalls and strategies in online experiments". *Information and Software Technology* 113, pp. 68–81. ISSN: 0950-5849. DOI: 10.1016/j.infsof.2019.05.004.

Kamalbasha and Eugster (2021). "Bayesian A/B testing for business decisions". *Data science–analytics and applications*. Springer, pp. 50–57.

Karrer, Shi, Bhole, Goldman, Palmer, Gelman, Konutgan, and Sun (2021). "Network experimentation at scale". *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3106–3116.

Katzir, Liberty, and Somekh (2012). "Framework and algorithms for network bucket testing". *Proceedings of the 21st international conference on World Wide Web*. WWW '12. New York, NY, USA: Association for Computing Machinery, pp. 1029–1036. ISBN: 978-1-4503-1229-5. DOI: 10.1145/2187836.2187974.

Kohavi, Longbotham, Sommerfield, and Henne (2009). "Controlled experiments on the web: survey and practical guide". *Data Mining and Knowledge Discovery* 18.1, pp. 140–181. ISSN: 1384-5810, 1573-756X. DOI: 10.1007/s10618-008-0114-1.

Kohavi, Tang, and Xu (2020). *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. https://experimentguide.com/. Cambridge: Cambridge University Press. ISBN: 978-1-108-72426-5. DOI: 10.1017/9781108653985.

Kontotasiou (2021). *The Guide to Ethical A/B Testing: The Missing Component of Your Optimization Program*. convert.com/blog/a-b-testing/ethical-ab-testing-guide/.

Koutra (2017). "Designing experiments on networks". PhD thesis. University of Southampton. 222 pp.

Kramer, Guillory, and Hancock (2014). "Experimental evidence of massive-scale emotional contagion through social networks". *Proceedings of the National Academy of Sciences* 111.24, pp. 8788–8790.

Leskovec, Lang, and Mahoney (2010). "Empirical comparison of algorithms for network community detection". *Proceedings of the 19th international conference on World wide web*. WWW '10. New York, NY, USA: Association for Computing Machinery, pp. 631–640. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772755.

Letham, Karrer, Ottoni, and Bakshy (2019). "Constrained Bayesian Optimization with Noisy Experiments". *Bayesian Anal.* 14.2, pp. 495–519. DOI: 10.1214/18-BA1110.

Li, Chu, Langford, and Schapire (2010). "A contextual-bandit approach to personalized news article recommendation". *Proceedings of the 19th international conference on World wide web*. WWW '10. New York, NY, USA: Association for Computing Machinery, pp. 661–670. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772758.

Liu, Sun, Varshney, and Xu (2019). "Large-scale online experimentation with quantile metrics". *arXiv preprint arXiv:1903.08762*.

Liu, Mandel, Brunskill, and Popovic (2014). "Trading Off Scientific Knowledge and User Learning with Multi-Armed Bandits". *EDM*.

Lomas, Forlizzi, Poonwala, Patel, Shodhan, Patel, Koedinger, and Brunskill (2016). "Interface Design Optimization as a Multi-Armed Bandit Problem". *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. New York, NY, USA: Association for Computing Machinery, pp. 4142–4153. ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858425.

Lux (2018). *Analyzing Experiment Outcomes: Beyond Average Treatment Effects - Uber Engineering Blog*. https://eng.uber.com/analyzing-experiment-outcomes/. (Accessed on 03/21/2022).

McFarland (2012). *Experiment!: Website conversion rate optimization with A/B and multivariate testing*. New Riders. 190 pp. ISBN: 978-0-13-304008-1.

McFarland, Pow, and Glick (2018). *Quasi Experimentation at Netflix*. https://netflixtechblog.com/quasi-experimentation-at-netflix-566b57d2e362.

Miroglio, Zeber, Kaye, and Weiss (2018). "The effect of ad blocking on user engagement with the web". *Proceedings of the 2018 world wide web conference*, pp. 813–821.

Mucha, Richardson, Macon, Porter, and Onnela (2010). "Community Structure in Time-Dependent, Multiscale, and Multiplex Networks". *Science* 328.5980, pp. 876–878. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1184819. arXiv: 0911.1824.

Nandy, Basu, Chatterjee, and Tu (2020). "A/B testing in dense large-scale networks: design and inference". *Advances in Neural Information Processing Systems* 33, pp. 2870–2880.

Newman (2006). "Modularity and community structure in networks". *Proceedings of the National Academy of Sciences* 103.23. Publisher: National Academy of Sciences Section: Physical Sciences, pp. 8577–8582. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0601602103.

Parker, Gilmour, and Schormans (2017). "Optimal design of experiments on connected units with application to social networks". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 66.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/rssc.12170, pp. 455–480. ISSN: 1467-9876. DOI: 10.1111/rssc.12170.

Parks, Aurisset, and Ramm (2017). *Innovating Faster on Personalization Algorithms at Netflix Using Interleaving*. https://netflixtechblog.com/interleaving-in-online-experiments-at-netflix-a04ee392ec55.

Pokhiko, Zhang, Kang, and Mays (2019). "D-optimal Design for Network A/B Testing". *Journal of Statistical Theory and Practice* 13.4, p. 61. ISSN: 1559-8608, 1559-8616. DOI: 10.1007/s42519-019-0058-3. arXiv: 1902.00482.

Prentice (1989). "Surrogate endpoints in clinical trials: definition and operational criteria". *Statistics in medicine* 8.4, pp. 431–440.

Radlinski and Craswell (2013). "Optimized Interleaving for Online Retrieval Evaluation". *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. WSDM '13. Rome, Italy: Association for Computing Machinery, pp. 245–254. ISBN: 9781450318693. DOI: 10.1145/2433396.2433429.

Rahman, Weiss, and Karunakaran (2023). "The Experimental Hand: How Platform-based Experimentation Reconfigures Worker Autonomy". *Academy of Management Journal* ja.

Rajkumar, Saint-Jacques, Bojinov, Brynjolfsson, and Aral (2022). "A causal test of the strength of weak ties". *Science* 377.6612, pp. 1304–1310.

Rudder (2014). *We experiment on human beings!* https://www.gwern.net/docs/psychology/okcupid/weexperimentonhumanbeings.html.

Sadeghi, Chien, and Arora (2019). "Sliced Designs for Multi-Platform Online Experiments". *Technometrics*. Publisher: Taylor & Francis, pp. 1–16. ISSN: 0040-1706. DOI: 10.1080/00401706.2019.1647288.

Saint-Jacques, Varshney, Simpson, and Xu (2019). "Using Ego-Clusters to Measure Network Effects at LinkedIn". *arXiv preprint arXiv:1903.08755*.

Sangho Yoon (2018). *Designing A/B tests in a collaboration network*. The Unofficial Google Data Science Blog. Library Catalog: www.unofficialgoogledatascience.com. URL: http://www.unofficialgoogledatascience.com/2018/01/designing-ab-tests-in-collaboration.html (visited on 06/11/2020).

Saveski, Pouget-Abadie, Saint-Jacques, Duan, Ghosh, Xu, and Airoldi (2017). "Detecting Network Effects: Randomizing Over Randomized Experiments". *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

KDD '17. Halifax, NS, Canada: Association for Computing Machinery, pp. 1027–1035. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098192.

Scott (2010). "A modern Bayesian look at the multi-armed bandit". *Applied Stochastic Models in Business and Industry* 26.6, pp. 639–658. ISSN: 1524-1904. DOI: 10.1002/asmb.874.

— (2015). "Multi-armed bandit experiments in the online service economy". *Applied Stochastic Models in Business and Industry* 31.1, pp. 37–45. ISSN: 1524-1904. DOI: 10.1002/asmb.2104.

Singer (2022). *LinkedIn Ran Social Experiments on 20 Million Users Over Five Years*. https://www.nytimes.com/2022/09/24/business/linkedin-social-experiments.html.

Stanley, Shai, Taylor, and Mucha (2016). "Clustering Network Layers with the Strata Multilayer Stochastic Block Model". *IEEE Transactions on Network Science and Engineering* 3.2. Conference Name: IEEE Transactions on Network Science and Engineering, pp. 95–105. ISSN: 2327-4697. DOI: 10.1109/TNSE.2016.2537545.

Stevens and Hagar (2022). "Comparative Probability Metrics: Using Posterior Probabilities to Account for Practical Equivalence in A/B tests". *The American Statistician* 76.3, pp. 224–237.

Stucchio (2015). "Bayesian A/B Testing at VWO", p. 33.

Ugander, Karrer, Backstrom, and Kleinberg (2013). "Graph cluster randomization: network exposure to multiple universes". *arXiv:1305.6979 [physics, stat]*. arXiv: 1305.6979.

Van Goffrier, Maystre, and Gilligan-Lee (2023). "Estimating long-term causal effects from short-term experiments and long-term observational data with unobserved confounding". *Proceedings of Machine Learning Research*. Vol. 213, pp. 786–794.

Walker and Muchnik (2014). "Design of Randomized Experiments in Networks". *Proceedings of the IEEE* 102.12. Conference Name: Proceedings of the IEEE, pp. 1940–1951. ISSN: 1558-2256. DOI: 10.1109/JPROC.2014.2363674.

Wildman (2019). *Using A/B Testing, Factorial Design, and Multivariate Tests for Deep Visitor Insights*. Library Catalog: www.thecreativemomentum.com. URL: https://www.thecreativemomentum.com/blog/using-a/b-testing-factorial-design-and-multivariate-tests-for-deep-visitor-insights (visited on 07/13/2020).

Xie and Aurisset (2016). "Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, pp. 645–654. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939733.

Zhang, Du, Andersen, and Hel (2022). *Beyond A/B Test: Speeding up Airbnb Search Ranking Experimentation through Interleaving*. https://medium.com/airbnb-engineering/beyond-a-b-test-speeding-up-airbnb-search-ranking-experimentation-through-interleaving-7087afa09c8e.

Zhang and Kang (2020). "Optimal Design for A/B Testing in the Presence of Covariates and Network Connection". *arXiv:2008.06476 [stat]*. arXiv: 2008.06476.

Zhou, Liu, Li, and Hu (2020). "Cluster-Adaptive Network A/B Testing: From Randomization to Estimation". *arXiv:2008.08648 [stat]*. arXiv: 2008.08648.