Interleaved Online Testing in Large-Scale Systems

Nan Bi*
Amazon, Palo Alto, CA, USA

Bai Li*
Amazon, Seattle, WA, USA

Ruoyuan Gao
Amazon, Palo Alto, CA, USA

Graham Edge
Amazon, Seattle, WA, USA

Sachin Ahuja
Amazon, Palo Alto, CA, USA

ABSTRACT

Online testing is indispensable in decision making for information retrieval systems. Interleaving emerges as an online testing method with orders of magnitude higher sensitivity than the pervading A/B testing. It merges the compared results into a single interleaved result to show to users, and attributes user actions back to the systems being tested. However, its pairwise design also brings practical challenges to real-world systems, in terms of effectively comparing multiple (more than two) systems and interpreting the magnitude of raw interleaving measurement. We present two novel methods to address these challenges that make interleaving practically applicable. The first method infers the ordering of multiple systems based on interleaving pairwise results with false discovery control. The second method estimates A/B effect size based on interleaving results using a weighted linear model that adjust for uncertainties of different measurements. We showcase the effectiveness of our methods in large-scale e-commerce experiments, reporting as many as 75 interleaving results, and provide extensive evaluations of their underlying assumptions.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**; • **Applied computing → Online shopping**.

KEYWORDS

Online evaluation, interleaved evaluation, A/B testing, e-commerce search

ACM Reference Format:
Nan Bi, Bai Li, Ruoyuan Gao, Graham Edge, and Sachin Ahuja. 2023. Interleaved Online Testing in Large-Scale Systems. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30-May 4, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3543873.3587572

1 INTRODUCTION

Online evaluation drives decision making for information retrieval (IR) and recommender systems [12, 13]. Typical IR system development follows an offline-online user feedback loop: formulate users' information need as an objective function, optimize the objective on historical user data, and finally evaluate the models on real users. Online evaluation is indispensable because there have always been gaps between offline formulation and user behavior in reality; and it becomes even more important to guide offline model tuning with the increased complexity in machine learning techniques.

A/B testing has been the predominant method for online evaluation. It shows candidate IR systems to randomized groups of users and compares user feedback metrics such as clicks or views. Despite being straightforward, A/B testing can slow down innovations as they require weeks and high user traffic to reach statistically significant conclusions. This happens when user metrics are noisy (e.g. clicks, views) and sparse (e.g. streams, purchases), and when IR systems grow mature with smaller-effect innovations.

Interleaving emerges as a more sensitive online testing method to free up experimentation bandwidth and expedite innovations [2, 4, 9, 11, 15–17]. Instead of presenting separate users with control and treatment results, interleaving merges their results into a single interleaved result and presents to all users. User actions on the interleaved result are attributed back to the two IR systems being compared, and the better one is whichever received (statistically significantly) more attributed actions. Literature report orders of magnitude higher sensitivity (10-100x) from interleaving than A/B testing when comparing which IR system is better [2, 4, 8, 10].

However, practical challenges have limited applicability of interleaving in IR and search systems. Since it is a paired test that directly evaluates user preference between two candidate systems, interleaving measures user feedback metrics in presence of both systems, which is not the same as A/B testing where absolute metrics are measured on each individual system. This means the raw results cannot directly order multiple (more than two) systems [4]. Methods have been proposed in dueling-bandits domain [19–21] that find the best candidate by minimizing regret; however, we would want to order all candidates up to certain confidence and select a subset for subsequent development or a follow-up A/B test for launching decisions. Moreover, the bandits methods require running multiple rounds of experiments that have practical difficulties in large-scale implementation. Another practical need is to interpret raw interleaving results in the magnitude of A/B measurements, to decide whether launching the new system worth the development cost. These magnitudes are also useful for applications that optimize and trade off multiple user metrics.

In this paper, we present two novel methods that address the aforementioned challenges to make interleaving practically applicable to IR system improvement. The first method infers ordering of multiple candidates based on interleaving pairwise results, with false discovery rate control under a hypothesis testing framework. We can now run interleaving to select a few top candidates from a large pool for a follow-up A/B test. The second method estimates the A/B effect size based on interleaving results using a weighted linear model that adjust for uncertainties of different measurements,

*Both authors contributed equally to this research.

WWW '23 Companion, April 30-May 4, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04.
https://doi.org/10.1145/3543873.3587572

so that history experiments with varying statistical confidence and power can all be utilized to improve the accuracy. Based on the estimated A/B effect size, we further introduce a power analysis to evaluate the statistical power required for the follow-up A/B test to detect the treatment effect of the selected systems. We will illustrate the effectiveness of our methods in real world e-commerce experiments at Amazon search, as well as provide theoretical and empirical evaluations of the assumptions behind our methods.

Contributions. To our best knowledge, our methods are novel in making inference on the ordering of multiple IR systems and estimating the practical effect size based on interleaving pairwise measurements. These methods improve interpretation of interleaving results and boost its applicability in real-world IR systems. We can rely less on the offline-online loop and expedite innovations by directly testing online with the highly sensitive interleaving. Furthermore, we report as many as 75 large-scale online experiments that apply interleaving to e-commerce search and showcase the practical benefits of our methods. Finally we evaluate the transitivity and linearity assumptions as proposed in literature with extensive theoretical and empirical analysis.

2 INTERLEAVING FOR SEARCH SYSTEMS

Interleaving is a paired test that evaluates user preference between two IR systems. Figure 1 illustrates the basic idea in the search ranking context. First, zip the ranking results from two compared systems into one combined list to present to all users. Then, attribute user engagement credit on the interleaved list back to the compared systems, and decide the winner that receives more credit (through a statistical test). [4, 8, 11].

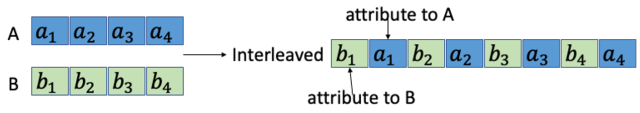

Figure 1: Ranking interleaving. Interleaving methods design algorithms to merge items and attribute user credits for whether two ranking lists overlap or not, to ensure fair comparison between two rankers.

Interleaving shows sensitivity boost ranging from 10 to 100 times over A/B testing in terms of required traffic to reach significance [2–4, 15], while preserving high fidelity to A/B conclusions. It resonates with perceptual test designs where users give binary responses of their preference which are more informative than absolute scores for each candidate. It also ensures a fully within-subject test that compare two systems on the same users, searches, and context, reducing heterogeneity variation in A/B testing design [16].

To compare multiple (more than two) ranking systems, transitivity property was introduced [4] to define consistency among multiple interleaving pairs: for any ranker triplet A > B > C, the interleaving effect sizes should follow $AC \geq \max\{AB, BC\}$. Methods have been proposed in dueling bandits framework that find the best ranker in multi-round experiments by minimizing regret [19–21], but it can be hard to apply them in practice. First, it is not straightforward to interpret the regret in business terms. Second, these methods run a lot of rounds (at least log of number of rankers) to minimize regret, while multi-round experiments require non-trivial infrastructure for automation and dynamic traffic allocation that are unbiased of treatment exposure from previous rounds. Extensions of interleaving such as multileaved methods [3, 18] can directly compare more than two rankers, but their implementation in large scale experiments is challenging [2]. Instead, in section 3 we propose to run data-efficient interleaving pairs among rankers and utilize transitivity to infer their ordering, so that an arbitrary number of top rankers can go to subsequent development.

Besides ordering multiple rankers, we want to interpret interleaving results in the magnitude of A/B measurements. This helps gauge the treatment effect size if launching the new ranker. It also puts different user feedback metrics in the same picture when optimizing for multiple metrics. Empirical studies report strong linear correlations between interleaving and corresponding A/B testing metrics [2, 4, 14]. In section 4 we exploit linearity property and fit a weighted least squares model to estimate A/B effect size based on interleaving measurement. Then we empirically evaluate linearity assumption by quantifying sign agreement probability between interleaving and A/B results.

3 COMPARING MULTIPLE SYSTEMS

Denote the outcome of an interleaving comparison as $X_{A,B}$ between two rankers A, B. Depending on the application, the analysis unit can be a query or a session, and the outcome can be binary (win or lose) or real-valued (the amount of attributed credit difference) [2]. Assume $X_{A,B}$ is i.i.d. with unknown distribution of mean μ_{AB} and variance σ^2_{AB}. Define ranker A is better than B ($A > B$) if $\mu_{AB} > 0$. We run a z-test on observations of $X_{A,B}$, and conclude ranker A is better than B if it is significantly positive. Define the transitivity property on true mean μ_{AB}.

Definition 3.1 (Transitivity Property of the Means (TPM))**.** For any triple A, B, C, with true ordering $A > B > C$, their session means satisfy that $\mu_{AC} \geq \max\{\mu_{AB}, \mu_{BC}\}$.

[TPM] says that interleaving results keep the magnitudes with respect to ordering of rankers. In Appendix A.1 we discuss the close connection between [TPM] and the original transitivity introduced in literature [20, 21]. In Appendix A.2 we prove that A/B testing results satisfy [TPM] by definition. Now we introduce a transitivity property [WTP] which is the fundamental assumption for our method that compare multiple rankers with interleaving results. [WTP] is implied by [TPM], see the proof in Appendix A.3.

Definition 3.2 (Weak Transitivity Property (WTP))**.** For any triple A, B, C, if $A > B$ and $B > C$, then it must be true that $A > C$.

Our method is as follows. Describe interleaving results in a graph. For K rankers and an arbitrary M of interleaving pairs among them, denote the rankers as nodes, and draw an edge pointing from ranker A to ranker B if there is an interleaving pair comparing them and ranker A is significantly better than B under confidence level α. Once finished with all the significant pairs, any ranker X is significantly better than any ranker Y if there is a directed path from node X to node Y.

This method is a direct application of [WTP]. One can treat each connected component in the graph as essentially a separate experiment, and apply Bonferroni correction [7] to control the overall

family-wise error rate for each connected component. Specifically, α can be the targeted overall error rate α_N divided by the number of interleaving pairs in the connected component. One can also consider Benjamini-Hochberg procedure for a less conservative control on false discovery rate [1].

This method can also evaluate [WTP] with experiment data. Specifically, we declare a violation to [WTP] at the overall confidence level α_N if and only if there exists a directed circle in the graph, which indicates interleaving results that are significant and directionally inconsistent. If there is no violation at α_N, then there is no violation at any more stringent confidence level $\alpha' < \alpha_N$.

4 TREATMENT EFFECT MAPPING

While transitivity is useful for comparing multiple rankers, it compares rankers qualitatively and ignores insignificant pairwise comparisons that might still carry useful information. On the other hand, A quantitative understanding of the treatment effect is often necessary in cases when a model is launched and we would like to learn the business impact, or when we optimize the trade-off between multiple metrics. Although interleaving provides quantitative measurements on users' preference, it cannot directly translate to business metrics measured in standard A/B experiments.

In this section, we provide a quantitative method for mapping the effect size measured in an interleaving experiment to the corresponding effect size in an A/B experiment that assesses the same treatment effect. We will refer to such A/B experiments as companion A/B experiments throughout the rest of the paper.

It has been shown in [2, 14], that there exists a strong linear correlation between the effect sizes in interleaving experiments and their companion A/B experiments. Assuming linearity, it is straightforward to consider fitting a linear model over the estimated effect sizes of interleaving experiments and their companion A/B. In practice, however, there exists heterogeneity among these data as they are measured under different statistical power. Hence, it is more appropriate to use the following weighted least square model:

$$\begin{aligned} \mathrm{ATE}^{AB} &= \mathrm{ATE}^{IL}\beta + \epsilon_i \\ \epsilon \sim N(0, \sigma^2/w) \quad & w = \mathrm{Var}(\mathrm{ATE}^{AB}) \end{aligned}$$

where ATE^{IL} is the average treatment effect (ATE) of interleaving experiments, ATE^{AB} is the average treatment effect of the companion A/B experiments and ϵ is the observation error. This formula weight data points inverse proportionally to their uncertainty. Note that we do not include the intercept coefficient for two reasons. First, when there is no treatment effect, we expect to see zero in both A/B and interleaving metrics. Second, both A/B and interleaving measurements should be symmetric w.r.t. the ordering of rankers, i.e. the treatment effect of ranker A versus B is negative of the effect of B versus A. These assumptions are solidified by using A/B testing results that have adjusted for pre-experiment biases [6] and that pairwise interleaving are immune to such bias by design.

Practically, we may fit this model to historical data and apply the mapping to any incoming interleaving measurements to estimate the corresponding A/B metrics, which then can be used for business decision making as in standard online testing schema.

4.1 Power Analysis

A common use case of interleaving is to select top rankers from a large candidate pool for a follow-up A/B experiment. With treatment effect mapping, we can calculate the power required for the follow-up A/B experiment to detect the treatment effect of these selected rankers by integrating over the distribution of the mapped treatment effect:

$$\int P(\text{detect the effect}|x) dN(x; \hat{\mathrm{ATE}}^{AB}, \mathrm{Var}(\hat{\mathrm{ATE}}^{AB}))$$

Here $P(\text{finding the effect}|x)$ is the power function of the A/B experiment given a constant effect size x, which is often available from standard power analysis. $N(x; \hat{\mathrm{ATE}}^{AB}, \mathrm{Var}(\hat{\mathrm{ATE}}^{AB}))$ is the distribution of the mapped treatment effect, where the mean is $\mathrm{ATE}^{IL}\mathbb{E}[\hat{\beta}]$, and the variance $\mathrm{Var}(\hat{\mathrm{ATE}}^{AB})$ is:

$$E(\mathrm{ATE}^{IL})^2\mathrm{Var}(\hat{\beta}) + \mathrm{Var}(\mathrm{ATE}^{IL})E(\hat{\beta})^2 + \mathrm{Var}(\hat{\beta})\mathrm{Var}(\mathrm{ATE}^{IL})$$

Here $\hat{\beta}$ is the estimated coefficient, whose mean and variance are easily obtainable.

4.2 Sign Disagreement

Although the treatment effect mapping provides a quantitative alignment between these interleaving and A/B, it cannot handle the case where there exists a sign disagreement between their estimated ATE. Notice such disagreement cannot be completely avoided due to the intrinsic uncertainty of experiments. Therefore, it is important to understand whether such disagreement is natural or due to unknown failures in interleaving. In this regard, we introduce summary statistics to track reliability of interleaving in terms of its sign disagreement with A/B experiments.

For any metric in an experiment, we assume the estimated average treatment effect has an approximate Gaussian distribution $N(\mu, \sigma^2/n)$ centered at the true treatment effect μ. Therefore, the probability of sign disagreement between ATE and μ is:

$$\begin{aligned} P(\mathrm{ATE} > 0|\mu < 0) =& P(\mathrm{ATE} < 0|\mu > 0) \\ =& \Phi(-n|\mu|/\sigma) \approx \Phi(-|\hat{\mathrm{ATE}}|/s) \end{aligned} \tag{1}$$

where s is the standard deviation of the estimated ATE. Clearly the probability of sign disagreement depends on the signal-to-noise ratio (SNR) $|\mathrm{ATE}|/s$. This aligns with the intuition that a noisy measurement is likely to cause sign disagreement.

For ease of analysis, we simplify the problem by assuming an interleaving experiment and its companion A/B experiment are merely two measurement on the same effect size μ with different powers. For this interleaving experiment and its companion A/B, the probability of sign disagreement is the probability that one and only one of them has a sign disagreement with the truth:

$$P(\mathrm{sign}(\mathrm{ATE}^{AB}) \neq \mathrm{sign}(\mathrm{ATE}^{IL})) = p_{AB} + p_{IL} - p_{AB}p_{IL}$$

where p_{AB} and p_{IL} represent the probability of sign disagreement between ATE and true treatment effect (equation 1) in A/B and interleaving respectively. With this formula, for any pair of interleaving and A/B experiments, we know the probability of observing a sign disagreement. Theoretically, the principle of hypothesis testing can be applied to determine whether a significant inconsistency exists between them. However, in practice, each pair of interleaving and A/B are not i.i.d. samples. Instead, we can monitor the empirical

distribution of the sign disagreements as summary statistics and determine whether it aligns with our theoretical computation. We explain this in details in Section 5.

5 ONLINE EXPERIMENT RESULTS

We present empirical results applying our methods to a range of e-commerce experiments at Amazon search. We follow the setup of [2] to use sessions as analysis unit and user feedback metrics such as clicks and purchases as real-valued outcome for interleaving comparisons. We demonstrate the effectiveness of our methods in real-world applications and also validate the assumptions of transitivity and linearity with these experiments.

Compare multiple ranking systems. We have run multiple interleaving experiments that efficiently compare rankers online, and observe no violations to transitivity property under the overall family wise error rate of 0.1. For illustration we show one experiment that evaluates 10 ranking systems with all 10-choose-2 interleaving pairs among them. We estimate that with classic A/B testing this would have taken more than 3x the time and 2.5x the traffic. Figure 2 plots the directed graph of the 10 rankers under overall confidence level of 0.1, and $\alpha = 0.1/45$ for each interleaving pair using Bonferroni correction. Here we trimmed some directed edges for easier visualization: Nodes in the same dashed box are those not significantly different from each other; and for any nodes X, Y that are not in the same box, a path from node X to node Y indicates $X > Y$. The overall inferred ordering is $C > \{R4, R6, R7, R8, R9\} > R5 > \{R2, R3\} > R1$. There are no directed loops so there is no violation of [WTP] detected at this confidence level.

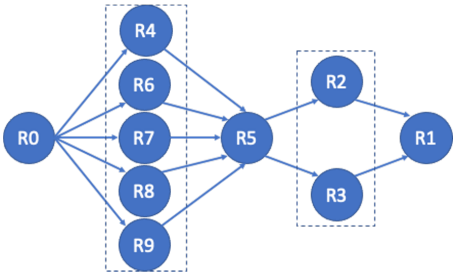

Figure 2: Compare multiple ranking systems. There are 10 rankers with directed edges for significant interleaving pairs. A path from X to Y indicates $X > Y$. Nodes in a dashed box are not significantly different from each other.

Treatment effect mapping. We use a dataset that consists of 75 pairs of interleaving measurements and their companion A/B measurements. Figure 3 plots two user feedback metrics, both of which show strong linearity between A/B and interleaving measurements up to their uncertainty. The regression lines with 95% confidence bands are also included. Most data points have their confidence intervals overlapped with the regression bands.

After obtaining the coefficients for metric 1 and metric 2, we can predict the effect size of a new treatment if tested in an A/B experiment. As an example, table 1 shows the estimated A/B effects for metric 1 and metric 2, and their actual A/B measurements. Although the estimated mean effects seem different from the actual

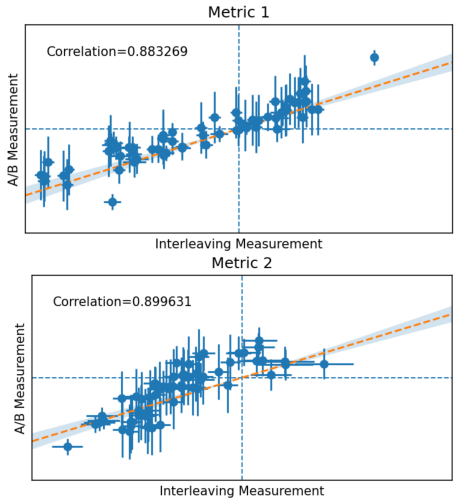

Figure 3: Treatment effect mapping for two user feedback metrics. Each datapoint shows the interleaving (x-axis) and the companion A/B measurement (y-axis) with corresponding 95% error bars. The orange dashed line is the fitted regression line with blue area indicating its 95% confidence band.

A/B effects, we believe this is due to the uncertainty within both the interleaving and A/B experiments. It is considered a good alignment as their confidence intervals are largely overlapped.

Table 1: Treatment effect mapping for a new treatment. The sample column indicates the percentage of samples used by interleaving versus A/B. The confidence intervals are largely overlapped, while interleaving requires much fewer samples.

Metrics	Estimated A/B effect	Actual A/B effect	Sample
Metric 1	-0.32% (-0.43%,-0.21%)	-0.15% (-0.42%, 0.12%)	1.15%
Metric 2	-0.07% (-0.20%,0.09%)	0.15% (-0.22%, 0.51%)	2.24%

Finally figure 4 plots a frequency histogram of sign disagreement in our dataset. We obtain 3 user metrics for each of the 75 experiment pairs. As a baseline, we also plot the expected probability of sign disagreement given the power of each experiment. Despite the variance, the overall trend matches the estimated probability, indicating the sign disagreement between A/B and interleaving measures are likely due to randomness.

This analysis can be used in various scenarios in practice. The most straightforward takeaway from equation 1 is that we need sufficient power in the experiments, no matter interleaving or A/B, to correctly measure the sign of the treatment effect. In addition, when seeing a disagreement between interleaving and its companion A/B experiment, we can calculate the probability of disagreement to explain whether such a disagreement is expected. Finally, we can maintain a set of interleaving and companion A/B pairs to continuously monitor the health of the interleaving system through a histogram similar to figure 4.

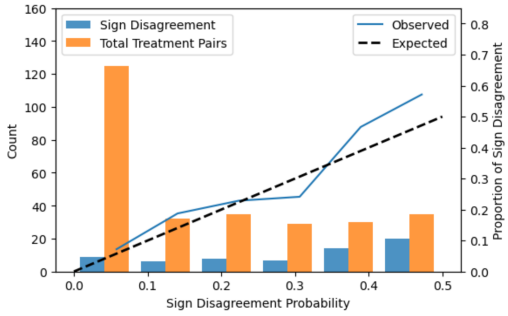

Figure 4: Observed sign disagreement proportions vs. expected probabilities. The histogram shows the numbers of total treatment pairs and the ones with sign disagreement. The blue line indicates the proportions of sign disagreement compared with the theoretical baseline in black dashed line.

6 CONCLUSION

We propose two novel methods to address the challenges of applying interleaving to real-world IR systems. The first method compares multiple systems with interleaving pairwise results while correctly controlling for false discovery rate. The second method estimates A/B effect size based on interleaving measurement, and utilize all history experiment results by taking their uncertainties into consideration. We demonstrate the applicability of these methods with 75 large-scale online experiment results, and further verify the underlying transitivity and linearity assumptions with extensive theoretical and empirical analysis. For future work, we will build on these methods and explore online learning-to-rank in a multiple-objective setting [5].

REFERENCES

[1] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[2] Nan Bi, Pablo Castells, Daniel Gilbert, Slava Galperin, Patrick Tardif, and Sachin Ahuja. 2022. Debiased balanced interleaving at Amazon Search. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2913–2922.

[3] Brian Brost, Ingemar J. Cox, Yevgeny Seldin, and Christina Lioma. 2016. An Improved Multileaving Algorithm for Online Ranker Evaluation. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR 2016)*. ACM, New York, NY, USA, 745–748. https://doi.org/10.1145/2911451.2914706

[4] Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. 2012. Large-Scale Validation and Analysis of Interleaved Search Evaluation. *ACM Transactions on Information Systems* 30, 1, Article 6 (March 2012). https://doi.org/10.1145/2094072.2094078

[5] Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. 2020. Differentiable expected hypervolume improvement for parallel multi-objective Bayesian optimization. *Advances in Neural Information Processing Systems* 33 (2020), 9851–9864.

[6] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 123–132.

[7] Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American statistical association* 56, 293 (1961), 52–64.

[8] Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. 2011. A Probabilistic Method for Inferring Preferences from Clicks. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (Glasgow, Scotland, UK) *(CIKM 2011)*. ACM, New York, NY, USA, 249–258. https://doi.org/10.1145/2063576.2063618

[9] Katja Hofmann, Shimon Whiteson, and Maarten De Rijke. 2013. Fidelity, Soundness, and Efficiency of Interleaved Comparison Methods. *ACM Transactions on Information Systems* 31, 4, Article 17 (Nov. 2013). https://doi.org/10.1145/2536736.2536737

[10] Kojiro Iizuka, Yoshifumi Seki, and Makoto P. Kato. 2021. Decomposition and Interleaving for Variance Reduction of Post-Click Metrics. In *Proceedings of the 7th ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR 2021)*. ACM, New York, NY, USA, 221–230. https://doi.org/10.1145/3471158.3472235

[11] Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. 2015. Generalized Team Draft Interleaving. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM 2015)*. ACM, New York, NY, USA, 773–782. https://doi.org/10.1145/2806416.2806477

[12] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online Controlled Experiments at Large Scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Chicago, Illinois, USA) *(KDD 2013)*. ACM, New York, NY, USA, 1168–1176. https://doi.org/10.1145/2487575.2488217

[13] 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. https://doi.org/10.1007/s10618-008-0114-1

[14] Joshua Parks, Juliette Aurisset, and Michael Ramm. 2017. Innovating Faster on Personalization Algorithms at Netflix Using Interleaving. hhttps://netflixtechblog.com/interleaving-in-online-experiments-at-netflix-a04ee392ec55

[15] Filip Radlinski and Nick Craswell. 2010. Comparing the Sensitivity of Information Retrieval Metrics. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Geneva, Switzerland) *(SIGIR 2010)*. ACM, New York, NY, USA, 667–674. https://doi.org/10.1145/1835449.1835560

[16] Filip Radlinski and Nick Craswell. 2013. Optimized Interleaving for Online Retrieval Evaluation. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining* (Rome, Italy) *(WSDM 2013)*. ACM, New York, NY, USA, 245–254. https://doi.org/10.1145/2433396.2433429

[17] Anne Schuth, Katja Hofmann, and Filip Radlinski. 2015. Predicting Search Satisfaction Metrics with Interleaved Comparisons. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR 2015)*. ACM, New York, NY, USA, 463–472. https://doi.org/10.1145/2766462.2767695

[18] Anne Schuth, Floor Sietsma, Shimon Whiteson, Damien Lefortier, and Maarten de Rijke. 2014. Multileaved Comparisons for Fast Online Evaluation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (Shanghai, China) *(CIKM 2014)*. ACM, New York, NY, USA, 71–80. https://doi.org/10.1145/2661829.2661952

[19] Yanan Sui, Masrour Zoghi, Katja Hofmann, and Yisong Yue. 2018. Advancements in Dueling Bandits.. In *IJCAI*. 5502–5510.

[20] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. 2012. The k-armed dueling bandits problem. *J. Comput. System Sci.* 78, 5 (2012), 1538–1556.

[21] Yisong Yue and Thorsten Joachims. 2011. Beat the mean bandit. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer, 241–248.

A APPENDIX

A.1 [TPM] and Transitivity in the Literature

We first review the formulation of transitivity property as introduced in the literature [20, 21]. Define a "duel" as an interleaving comparison between two rankers for a query, with a binary outcome of which ranker wins this query. The duels are i.i.d. according to a $K \times K$ preference matrix P, where $P_{i,j}$ is the probability that ranker i beats ranker j, i.e. each duel of i and j is a Bernoulli draw of $P_{i,j}$. Denote the winning margin of i over j by $\Delta_{i,j} = P_{i,j} - 1/2$. Ranker i is better than j (i.e. $i > j$) if $P_{i,j} > 1/2$ or equivalently $\Delta_{i,j} > 0$. The transitivity property is defined as below that says the probability of winning a duel keeps the ordering with respect to multiple rankers.

Definition A.1 (Strong Stochastic Property (SST))**.** For any triple (i, j, k), with true ordering $i > j > k$, the preference matrix satisfies that $\Delta_{i,k} \geq \max\{\Delta_{i,j}, \Delta_{j,k}\}$.

Our [TPM] is closely related to [SST]. It is an extension so that the outcome of interleaving comparison can describe the magnitude

of metrics such as revenue. Suppose we consider ranker A beats B in their duel if it wins directionally $\bar{X}_{A,B} > 0$, and correspondingly $P_{A,B} = P(\bar{X}_{A,B} > 0)$ in the preference matrix P. [TPM] will imply [SST] if the sample sizes and variances are approximately equal $N_{AC} = N_{AB} = N_{BC}, \sigma^2_{AC} = \sigma^2_{AB} = \sigma^2_{BC}$. To see this, consider the definition of Z statistic $Z_{AB} = \frac{\bar{X}_{AB}-0}{s_{AB}}$, where the null hypothesis is $\mu_0 = 0$, and standard error $s_{AB} = \frac{\hat{\sigma}_{AB}}{\sqrt{N_{AB}}}$. The distribution of Z statistic under the true mean μ_{AB} is that

$$Z_{AB} = \frac{\bar{X}_{AB} - \mu_{AB}}{s_{AB}} + \frac{\mu_{AB}}{s_{AB}} \approx N(0,1) + \sqrt{N_{AB}}\frac{\mu_{AB}}{\sigma_{AB}} = N\left(\sqrt{N_{AB}}\frac{\mu_{AB}}{\sigma_{AB}}, 1\right)$$

The probability of A winning over B is

$$P(\bar{X}_{AB} > 0) \approx P\left(N\left(\sqrt{N_{AB}}\frac{\mu_{AB}}{\sigma_{AB}}, 1\right) > 0\right)$$

Assume large sample size N, equal or close sample sizes and variances, and all positive means μ, the winning probability increases with the magnitude of μ, hence satisfies [SST].

A.2 A/B Testing Satisfies [TPM]

[TPM] setup for classic A/B testing. Assume a ranker A's session metric X_A is i.i.d. with unknown distribution of mean μ_A and variance σ^2_A. For a triplet of independent rankers A, B, C such that $A > B > C$, the mean of the two sample T statistic for rankers A and B is $\mu_{AB} = \mu_A - \mu_B$, and similar for B, C and A, C. Then because $\mu_{AB}, \mu_{BC}, \mu_{AC} > 0$, we have

$$\mu_{AC} = \mu_{AB} + \mu_{BC} \geq \max\{\mu_{AB}, \mu_{BC}\}$$

A.3 [TPM] Implies [WTP]

We know that $\mu_{AB}, \mu_{BC} > 0$ from the fact that $A > B$ and $B > C$. [TPM] tells us that $\mu_{AC} \geq \max\{\mu_{AB}, \mu_{BC}\} > 0$, which means $A > C$.