Debiased Balanced Interleaving at Amazon Search

Nan Bi
Amazon, Palo Alto, CA, USA

Pablo Castells
Amazon, Madrid, Spain

Daniel Gilbert
Amazon, Palo Alto, CA, USA

Slava Galperin
Amazon, Palo Alto, CA, USA

Patrick Tardif
Amazon, Palo Alto, CA, USA

Sachin Ahuja
Amazon, Palo Alto, CA, USA

ABSTRACT

Interleaving is an online evaluation technique that has shown to be orders of magnitude more sensitive than traditional A/B tests. It presents users with a single merged result of the compared rankings and then attributes user actions back to the evaluated rankers. Different interleaving methods in the literature have their advantages and limitations with respect to unbiasedness, sensitivity, preservation of user experience, and implementation and computation complexity. We propose a new interleaving method that utilizes a counterfactual evaluation framework for credit attribution while sticking to the simple ranking merge policy of balanced interleaving, and formally derive an unbiased estimator for comparing rankers with theoretical guarantees. We then confirm the effectiveness of our method with both synthetic and real experiments. We also discuss practical considerations of bringing different interleaving methods from the literature into a large-scale experiment, and show that our method achieves a favorable tradeoff in implementation and computation complexity while preserving statistical power and reliability. We have successfully implemented our method and produced consistent conclusions at the scale of billions of search queries. We report 10 online experiments that apply our method to e-commerce search, and observe a 60x sensitivity gain over A/B tests. We also find high correlations between our proposed estimator and corresponding A/B metrics, which helps interpret interleaving results in the magnitude of A/B measurements.

CCS CONCEPTS

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

KEYWORDS

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

ACM Reference Format:
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 and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3511808.3557123

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557123

1 INTRODUCTION

Online evaluation is the yardstick for improvement and innovation of search systems [14,15,26]. A/B testing is the pervading online evaluation approach, where new search ranker candidates ("treatments") and the current ranker in production ("control") are presented to randomized groups of real users, and user feedback metrics such as clicks and purchases are used to compare and determine the winning variant.

A/B tests can often be a bottleneck to search ranking updates and innovations as they typically take several weeks to reach statistically significant conclusions, and soon become a scarce resource as demand for experimentation outgrows the available bandwidth of online traffic. Reasons for lack of sensitivity include user feedback being noisy (e.g. clicks) and/or sparse (e.g. purchases), and effect sizes of new innovations tend to shrink as the search system grows mature. Moreover, during an A/B test users are exposed to potentially suboptimal rankers – the longer this goes on, the wider the impact of undesired effects in their search experience.

In need of improving the sensitivity of online evaluation, interleaving took shape as a new technique and was gradually adopted by industry [1,3,6,7,10,19,21,25]. Rather than showing the evaluated rankers separately, interleaving presents users with a single merged list of the compared rankings, observes and attributes user actions on the combined ranking back to the evaluated rankers. The advantage of interleaving is that user reaction to results from one system can be observed *in the presence* of results from other competing systems. This enables relative preference measurements that are not possible when testing rankers separately. Empirical studies have demonstrated orders of magnitude higher sensitivity of interleaving than traditional A/B tests [3,6,8,9].

Several evaluation criteria for a good interleaving method have been proposed in the literature: correctness of the declared winning ranker, sensitivity to improvements in search quality, and preserving user experience quality as an online intervention [6,7,13,20]. Interleaving methods distinguish from each other by how the interleaved result is constructed from the compared rankings (interleaving policy), and how user actions are attributed back to the compared rankers (credit attribution) [3]. Each method has its own advantages and limitations with respect to these criteria. The simplest interleaving method, *balanced interleaving* [10,11], is biased when comparing lists that are similar but up to small shifts in position [7,21]. More elaborate interleaving methods are proposed to avoid this issue, at the expense of limitations in sensitivity, potential degradation of user experience, and/or increased implementation complexity and computational cost [3,6,7,13,17,20,23,24].

In this paper we propose a new approach that utilizes a counterfactual evaluation framework for credit attribution while sticking to the simple merging policy of balanced interleaving. The key idea in the proposed approach is to apply a propensity-based weight-

ing to the attributed credits, and formally derive a closed-form unbiased estimator for comparing rankers. To our best knowledge, the proposed method is novel with respect to state of the art in interleaved online evaluation. Our method has been implemented successfully and produces trustworthy conclusions at the scale of billions of search queries. We illustrate the effect of our method in both simulations and online experiments at Amazon search, showing improvement over alternative state-of-the-art interleaving methods with respect to the aforementioned evaluation criteria.

Our main contributions and findings include the following. First, the proposed method corrects the bias of balanced interleaving under theoretical guarantees while inheriting its merits of simplicity and sensitivity. We discuss the practical considerations for implementing these interleaving methods in large-scale experiments, and show that our method achieves a better tradeoff than other alternatives in terms of implementation and computational cost. Second, we report 10 large-scale online experiments applying interleaving to e-commerce search, taking sessions rather than queries as the analysis unit, and performance metrics based on purchases rather than clicks. We observe in this setting that balanced interleaving reaches the wrong conclusion in 3 out of 10 experiments while our proposed method corrects all of them. To our best knowledge, such systematic impact of the bias has not been reported before in an industry-scale search engine. As an additional finding, we report high correlations between our proposed estimator and corresponding business metrics in traditional (non-interleaved) A/B tests, which in practice helps interpret the interleaving credit values in the magnitude of A/B effect sizes.

2 INTERLEAVING IN ONLINE EVALUATION

Figure 1 (a) illustrates the idea of interleaving. A merged ranking is created by zipping two lists together, alternating turns to pick elements from the lists into the combined ranking. When the interleaved list is presented to users, the ranker that generates more user engagement through its elements in the merged list wins the comparison. An interleaving method thus comprises a) an interleaving protocol for sorting turns when selecting elements from the compared rankings and placing them in the merged list, and b) a credit attribution policy to decide which ranker gets credit for user actions on elements in the merged ranking.

Interleaving achieves high sensitivity due to its unique design. When choosing one among several alternatives, direct relative measurements between systems are more informative than absolute observations [16]. Moreover, interleaving turns online tests into a fully within-subject design, where the competing rankers are exposed to the exact same users and the same searches, removing the noise in assigning different treatments to different users [20].

2.1 Evaluation Criteria

We describe the desirable properties of a successful interleaving method, following requirements that have been proposed in the literature [6,7,13,20]. We will use these evaluation criteria when reviewing different interleaving methods and motivating our proposed method in sections that follow.

- Unbiasedness: the method should be unbiased, i.e. it should not prefer either ranker in expectation when user engagement with the interleaved rankings is random.
- Fidelity: the method displays fidelity if the expected outcome aligns with the true comparison between rankers. The direction of preference should agree with the ground truth.
- Sensitivity: the method should be able to detect small differences ranker quality. It should be data-efficient, reaching statistically significant conclusions with as few observations as possible.
- User experience quality: the method should not substantially alter or degrade the search experience of real users during the experiments.
- Scalability: the method should be affordable in large-scale experiments for implementation complexity and computation cost.

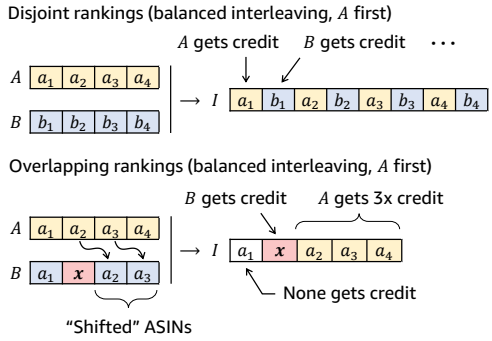

Figure 1: Ranking interleaving.

The first three criteria define the correctness and effectiveness of the method. The conclusions should align with true comparisons, both when there is a true difference between rankers, and when there is not. The last two criteria highlight practical considerations for the method to be widely adopted in industry. The method must preserve the user experience quality as an online intervention; otherwise users may quickly abandon a search system that performs poorly.

2.2 Interleaving Methods

Balanced interleaving is the simplest and most inexpensive interleaving approach [10,11]. It starts with a coin toss to decide which ranker goes first. Then rankers take turns to pick their next element from top to bottom, and place them in the interleaved ranking until one of the lists is exhausted. A ranker simply misses the turn if its picked element is already in the interleaved ranking. A cutoff position is defined for credit attribution as the highest position that, when applied to rankings A and B, produces the interleaved ranking down to the last user action. User actions are credited to one or both rankers if they rank the element above the cutoff position. Alternative attribution policies are suggested in the literature [3] – regardless, balanced interleaving suffers from systematic bias in specific edge cases [7], as we discuss in detail in the next subsection.

Team draft interleaving [21] avoids the bias in balanced interleaving by tossing a coin not just at the beginning, but at every turn, to decide which ranker goes first. In their respective turn, rankers place their next item down the list that is not yet present in the interleaved ranking until one ranker reaches its end position. User actions are credited to the ranker that happened to select the element. Team draft has its own breaking cases where it fails to detect the difference between certain input rankings [3,7]. We illustrate this weakness through simulations in Section 6.

This limitation is overcome by *probabilistic* interleaving [6,7], that allows any element other than just the next one to be picked, where selection probability decreases with ranking depth. Credit attribution is considerably more involved than in balanced and team

draft interleaving: the probabilities of all possible combinations of coin tosses and element samplings are computed that can lead to the produced interleaved ranking. For each such sequence of steps, the ranker that would have placed each element in the interleaved list is credited for the actions upon that element; the actions are then weighted by the probability of the sequence and the whole is summed into credit expectation for each ranker [6,7]. Downsides of probabilistic interleaving are notably the potential degradation of the user experience [17], and the prohibitive computation cost [6].

Further research in this area has explored interleaving extensions such as multileaved methods that compare more than two rankers [1,9,23,25], and sophisticated techniques that learn to optimize the interleaving policy and credit attribution to meet desired properties and constraints [9,12,20,25,27]. Given the increased complexity involved in these methods, practical challenges arise when envisioning large-scale experiments, where simpler methods may present preferable tradeoffs, as we discuss later in Section 3.

2.3 Overlapping Imbalance

Balanced interleaving introduces systematic bias when the input rankings overlap and one or more common elements are placed higher by one of the rankers. We refer to this issue as *overlapping imbalance*. Bias here means that one ranker is found better than the other in expectation when exposed to randomly behaving users [7].

Figure 1 (b) illustrates the overlapping imbalance with a toy example. Here we take a credit policy that attributes user actions on common elements only to the ranker that placed the element higher – other credit attribution variants generate a similar bias, as described in the literature [3]. The two rankings are $A = \langle a_1, a_2, a_3, a_4 \rangle$, $B = \langle a_1, x, a_2, a_3 \rangle$. By injecting x in the ranking, B shifts a_2 and a_3 down compared to A. The interleaved list can be, with equal probability, $I = \langle a_1, x, a_2, a_3, a_4 \rangle$ or $\langle a_1, a_2, x, a_3, a_4 \rangle$, depending on which ranker starts first. By definition, a randomly behaving user acts on each position of I with equal probability, and should produce no preference about either ranker. However, since B ranks x higher than A, and A ranks a_2, a_3, a_4 higher than B, actions on x are credited to B while a_2, a_3, a_4 are attributed to A. This creates a 3:1 a priori advantage for A over B; as a result, A is expected to get three times as much credit for user actions as B under random behavior.

This toy example is representative of the general case where common elements of the two lists are systematically ranked higher in one of them, giving this list an unfair a priori advantage to receive credit. Contrary to the expectation that these edge cases are infrequent or at worst evenly affect the competing systems in randomized experiments [3], we find them not rare in Amazon search experiments – we observe bias against one particular side of the rankers, to the extent of reversing the sign of comparisons. For instance, a treatment that pushes some item up in the control ranking will effectively shift all elements below this item down by one position, putting itself at a systematic disadvantage compared to the control ranker. This systematic imbalance motivates the solution we propose in this paper.

3 LARGE-SCALE INTERLEAVING AT AMAZON SEARCH

Our proposed method retains the simplicity of the original balanced interleaving, which we find to be critical for large-scale experiments in terms of implementation complexity and computational cost. There are wider implications in infrastructure and engineering beyond merely the implementation of the core algorithm. For online impact in a large-scale search engine, pre-computing and caching search results of frequent queries is a critical optimization for query-time latency: almost one third of the Amazon search traffic is served out of cached results. With our method or balanced interleaving, there are only two possible results for a given query depending on the coin flip. With team-draft there can be as many as 2^K different possible merged rankings for a given query, where K is the size of the interleaved ranking – this essentially reduces the caching efficiency by a factor of 2^K. For probabilistic interleaving, any permutation is also possible in the merged ranking, so the efficiency penalty is even worse, by factorial $K!$

Moreover, the extra ranking randomness in team draft and probabilistic interleaving requires a user-level caching mechanism, unless one is willing to have a changing ranking every time a user refreshes the page or enters the same query twice. Furthermore, team draft requires logging all the coin tosses to compute credit attribution in the post-experiment analysis. Such mechanisms are conceptually simple, but arguably a substantial burden in engineering cost and technical debt. Another major concern with probabilistic interleaving is the alteration of the original order of the evaluated rankers. The resulting degradation of search quality puts user trust at risk, especially when it happens systematically at scale [18].

Finally, in post-experiment analysis, exact credit computation for probabilistic interleaving has exponential complexity; even with Monte Carlo approximation, the cost is a multiple of balanced interleaving, linear in the number of Monte Carlo samples [7,17,18, 23]. This can be a hindrance to producing robust and precise results at the scale of billions of queries and hundreds of positions.

Optimized interleaving methods [9,12,20,25,27] have also shown to effectively achieve the desirable properties of an interleaved approach. They involve however similar or greater complexity issues than discussed above, which may not find an easy justification given the effectiveness of our proposed alternative. Compared to optimized interleaving with ad hoc rules and history data, our method is more robust to search system changes and user behavior shifts. As for scaling experiments to higher numbers of rankers, it is straightforward to generalize our method to multileaved comparisons [1,9,23,25], as e.g. team draft is generalized in [25]. However, the cost of answering one query gets multiplied by the number of rankers, which linearly degrades the query latency to users. We instead envision running multiple data-efficient pairwise interleaved comparisons within sets of rankers as a better alternative, relying on transitivity to discard suboptimal rankers and select the best ones with fair confidence [3].

4 DEBIASED BALANCED INTERLEAVING

We address the overlapping imbalance described in Section 2.3 by a counterfactual evaluation approach applied to balanced interleaving – specifically, a form of inverse propensity weighting (IPW) [22]. In our proposed approach, the attributed credit to a given ranker for an action in the interleaved ranking is normalized by the propensity (the probability) of the competing ranker to be attributed credit at that position. In IPW terminology, each position k in an interleaved impression can be seen as an "individual" to whom a treatment A or B is to be dispensed, and credit (customer engagement) represents

the outcome of interest of the treatment. Credit attribution to one ranker at an interleaved position k is equivalent, in this view, to assigning the ranker treatment to position k and observing whether or not customers engage with that position. We thus weigh the observed credit to a ranker by the inverse probability that the interleaved position would have been attributed to that ranker.

4.1 Formalization

Formally, let I denote an interleaved list of A and B. Let $\alpha_k(I) \in \{A, B, \emptyset\}$ denote the credit attribution decision at position k in I. We can express the credit attributed to a ranker $R \in \{A, B\}$ from an impression of an interleaved ranking I as:

$$C_R(I) = \sum_k \mathbb{1}(\alpha_k(I) = R)\, V_k(I)\, E_k(I)$$

where $V_k(I)$ is a binary variable equal to 1 if the user views the item at position k in I, $E_k(I)$ is the user engagement at position k, and $\mathbb{1}(X)$ is the binary indicator function (1 if X is true). While the literature has focused on binary user actions and binary ranker comparisons, user engagement $E_k(I)$ in our formulation can be binary (e.g. clicked / not clicked) or numeric (e.g. purchased units, revenue) [26]. Also we consider numeric experiment outcomes comparing rankers by the difference in attributed credit $C_A(I) - C_B(I)$.

Our debiased approach introduces the following normalization in the above equation:

$$\bar{C}_R(I) = \sum_k \frac{\mathbb{1}(\alpha_k(I) = R)\, V_k(I)}{P(\alpha_k(I) = R \mid \alpha_k(I) \neq \emptyset, V_k(I) = 1)}\, E_k(I) \qquad (1)$$

where the condition $\alpha_k(I) \neq \emptyset$ is needed for a well-defined probability distribution on $R \in \{A, B\}$. The normalization is essentially the propensity to be attributed position k of the interleaved list I.

When there is no overlapping imbalance, the normalizing factor is $1/2$ for both rankers, and the bias-corrected method becomes equivalent to uncorrected attribution. When there is a systematic overlapping imbalance, this propensity factor gets away from $1/2$ and helps restore the balance in the ranker comparison.

4.2 Attribution Propensity Estimation

To estimate the propensity, we make a simplifying assumption that ranker attribution α is independent from the position k such that $P(\alpha_k(I) = R \mid \alpha_k(I) \neq \emptyset) \approx P(\alpha(I) = R \mid \alpha(I) \neq \emptyset)$. This helps meet the IPW positivity condition [22] of a nonzero propensity: a ranker may have zero chance of getting credit for a specific position, but always gets nonzero attribution over the entire interleaved list. We can then estimate the propensity as the ratio of positions attributed to R over total nonzero attributed positions:

$$\widehat{P}(\alpha(I) = R \mid \alpha(I) \neq \emptyset) = \textstyle\sum_k \mathbb{1}(\alpha_k(I) = R) \;/\; \sum_k \mathbb{1}(\alpha_k(I) \neq \emptyset)$$

Plugging this into equation 1 we get our debiased estimator:

$$\bar{C}_R(I) = \frac{\sum_k \mathbb{1}(\alpha_k(I) = R)\, V_k(I)\, E_k(I)}{\sum_k \mathbb{1}(\alpha_k(I) = R)\, V_k(I) \;/\; \sum_k \mathbb{1}(\alpha_k(I) \neq \emptyset)\, V_k(I)} \qquad (2)$$

The correction can be interpreted as dividing the credit that each ranker R obtains by the ratio of total prior attribution to R over the impressed interleaved ranking I.

To illustrate the effect of our approach, consider again the toy example $A = \langle a_1, a_2, a_3, a_4 \rangle$ and $B = \langle a_1, x, a_2, a_3 \rangle$ from Section 2.3. As explained earlier, the number of positions attributed to A and B are $\sum_k \mathbb{1}(\alpha_k(I) = A) = 3$, and $\sum_k \mathbb{1}(\alpha_k(I) = B) = 1$. Per equation 2, credit to A should be divided by $3/(3+1)$ and credit to B by $1/(3+1)$. For a randomly behaving user with uniform engagement probability $P(E_k(I)) = p$, the corrected credit is $\mathbb{E}[\bar{C}_A(I)] = 4p = \mathbb{E}[\bar{C}_B(I)]$ in expectation. The corrected measurements are no longer biased since A earns as much credit (and wins as many times) as B in expectation. It is easy to check that this remains a tie even if the user browses I with non-uniform position probability. The unbiasedness proof in the next subsection reflects this generalization.

In our application, we group measurements and correction at a higher unit of sessions rather than individual impressions, as in the experiments reported later in Section 7. Sessions can be more meaningful than individual queries as a unit of randomization and comparison, as they comprehend a more complete account of a search intent [21]. To this end we make an additional independence assumption over impressions in a session S such that $P(\alpha(I) = R \mid \alpha(I) \neq \emptyset, I \in S) \approx P(\alpha(S) = R \mid \alpha(S) \neq \emptyset)$, and estimate the propensity by averaging over all the impressions of each session in an experiment Ω:

$$\bar{C}_R(\Omega) = \frac{1}{|\Omega|} \sum_{S \in \Omega} \frac{1}{O_R(S)} \sum_{I \in S} \sum_k \mathbb{1}(\alpha_k(I) = R)\, V_k(I)\, E_k(I) \qquad (3)$$

where:

$$O_R(S) = \sum_{I \in S} \sum_k \mathbb{1}(\alpha_k(I) = R)\, V_k(I) \;/\; \sum_{I \in S} \sum_k \mathbb{1}(\alpha_k(I) \neq \emptyset)\, V_k(I)$$

is the total credit opportunity that ranker R gets across session S. Finally the outcome of experiment Ω is defined as $\bar{C}_A(\Omega) - \bar{C}_B(\Omega)$, and it is straightforward to apply hypothesis testing (z-test) to produce p-values and confidence intervals.

4.3 Unbiasedness and Fidelity

We now demonstrate the correctness of our method with theoretical guarantees, specifically that our method satisfies unbiasedness and fidelity criteria [7,13,20] defined in section 2.1.

Unbiasedness means that a randomly behaving user should not create a preference for either ranker: we need to prove that the expected credit difference (as per equation 1) between two rankers is always zero in that case. A random user engages with an item at an examined position with identical distribution regardless of what the item is: $E_k(I) \approx E_k | V_k = 1$ and $V_k(I) \approx V_k$. Credit expectation is taken with respect to user engagement E, viewing status V, and credit attribution by the attribution policy α (V, E and α being random vectors over positions k in I).

Theorem 1 (Unbiasedness). *For a randomly behaving user with $E_k(I) \approx E_k \mid V_k = 1$ and $V_k(I) \approx V_k$, the attributed credit by equation 1 returns a tie in expectation:*

$$\mathbb{E}_{\alpha,V,E}[\bar{C}_A(I) - \bar{C}_B(I)] = 0$$

Proof. Starting from equation 1, and knowing that attribution α_k is independent from examination V_k and engagement E_k under random user behavior, we have:

$$\begin{aligned}
\mathbb{E}[\bar{C}_R(I)] &= \mathbb{E}_{\alpha,V,E}\left[\sum_k \frac{\mathbb{1}(\alpha_k(I) = R)\, V_k(I)}{P(\alpha_k(I) = R \mid \alpha_k(I) \neq \emptyset, V_k(I) = 1)}\, E_k(I)\right] \\
&= \sum_k \frac{\mathbb{E}_{\alpha,V}\left[\mathbb{1}(\alpha_k(I) = R)\, V_k(I)\right]\, \mathbb{E}_E[E_k(I) \mid V_k(I) = 1]}{P(\alpha_k(I) = R, V_k(I) = 1) \,/\, P(\alpha_k(I) \neq \emptyset, V_k(I) = 1)} \\
&= \sum_k \frac{P(\alpha_k(I) = R, V_k(I) = 1)\, \mathbb{E}_E[E_k(I) \mid V_k(I) = 1]}{P(\alpha_k(I) = R, V_k(I) = 1) \,/\, P(\alpha_k(I) \neq \emptyset, V_k(I) = 1)}
\end{aligned}$$

$$= \sum_k P(\alpha_k(I) \neq \emptyset, V_k = 1)\, \mathbb{E}_E[E_k \mid V_k = 1]$$

Since expected credit no longer depends on the ranker R, then for any two rankers A, B, we have that $\mathbb{E}[\bar{C}_A(I)] = \mathbb{E}[\bar{C}_B(I)]$ and the credit difference is zero in expectation. □

For fidelity, Hofmann et al. [7] propose that an interleaving method should declare (in expectation) a winner any ranking A that *Pareto-dominates* another ranking B. A Pareto-dominates B means that any relevant item contained in B is also contained and ranked higher in A. Radlinski et al. [20] propose that if a document d is clicked, the ranker that ranked d higher should be given more credit for the click than the other ranker. Note that we choose to use the attribution policy that attributes credit for common items only to the ranker that places it higher, and to neither if the item is ranked in the same position by both. This was one of the variants proposed in [3] for improved sensitivity over the original balanced interleaving. It can be easily seen that both properties are satisfied by our method, because user actions on items ranked higher in A are entirely credited to A, and none to B.

In comparison, balanced interleaving and team draft have known breaking cases that fail some of these requirements. Balanced interleaving does not satisfy the unbiasedness requirement. When there is an overlapping imbalance, balanced interleaving declares a wrong winner for random users [3]. Team-draft does not satisfy the fidelity requirement. The breaking cases are of the form $A = \langle a, b, x\rangle$, $B = \langle b, x, a\rangle$ where x is the only item engaged upon; team draft would declare a tie while the true winner is B since it ranks the relevant item x higher [7,20].

Our method has its own breaking cases where it might declare the wrong winner. These occur when a ranker "wastes" its attribution advantage given by an overlapping imbalance: if the shifted items attract scarce user engagement, the ranker is still penalized for an advantage it does not actually leverage. For instance, with $A = \langle x, a, b, y\rangle$, $B = \langle c, y, a, b\rangle$, x and y being the only engaged items, our method declares B a winner, even though A is better both in number and overall rank of the engaging items – A does not take advantage of being attributed a and b (these items get no engagement) while being penalized for it. However, contrary to the overlapping imbalance, the edge cases of our method are peculiar enough that we see no reason to expect a noticeable incidence in randomized experiments, even less to have an asymmetric effect.

An interleaving method with good fidelity should generally agree with other appropriate evaluation methods in identifying the better ranker [20]. To complement the above theoretical specifications, in Section 7 we will test the fidelity of our method with real search traffic by comparing with A/B test results. We will also show in Section 6.3 that our method is more effective than probabilistic interleaving in addressing the team draft edge case.

5 EMPIRICAL EVALUATION

We demonstrate the effect of our method with both simulations in Section 6 and real online experiments in Section 7. Simulations with synthetic rankings and user behavior allow for quick and flexible testing of fundamental properties and comparisons with alternatives [5]. The online experiments provide proof-of-concept for our approach when applied to real search traffic, and show the effectiveness of our method in large-scale experiments.

Together, our experiments on simulated and real search traffic aim to answer the following questions:

RQ1 Are interleaved comparisons unbiased? When user behavior is random, an unbiased experiment should declare a tie by not finding any statistically significant difference.

RQ2 How sensitive are the comparisons? If one ranker systematically places highly-engaging items higher than another ranker, the method should not miss the difference.

RQ3 Our method is defined on a particular credit attribution policy. Would it be effective on other attribution policy variants?

RQ4 For interleaved comparisons that return numeric outcomes, are they meaningful in assessing quantitative differences between rankers? How do the numeric interleaved outcomes correlate with A/B testing measurements?

6 SIMULATED EXPERIMENTS

The simulations compare our approach to other interleaving methods and variants.[1] The interleaving methods we consider include balanced interleaving [10,11], team draft [21], and probabilistic interleaving [6,7]. For the latter we use the implementation by Oosterhuis et al. [17], with the default setting $\tau = 0.3$, and 100 samples (we found no substantial improvement by further increasing the sample size). For balanced interleaving, we consider three credit attribution policies, suggested in [3]:

- *Original policy:* typically the default option in literature, credit on common items is attributed to both rankers down to a cutoff position defined by the lowest interaction [3,7,10,19,21].
- *Uncorrected policy:* credit is attributed to the ranker that ranks the item higher, and none if the positions are the same [3].
- *Weighted policy:* user actions are attributed to both rankers, weighted by 1/log of item position in the input ranking [3].

Along with the interleaved comparisons, we simulate a standard A/B test where the same rankers A and B are separately exposed to the simulated users, getting credit from standalone sessions.

6.1 Simulation setup

Similarly to prior work [5,7], we run synthetic simulations to evaluate the interleaved methods against the success criteria, with a focus on cases that our method aims to address. While we report here results on simple user models and ranking policies, we observe the same trends over a wider model variety we have tested.

6.1.1 Rankings. We generate artificial rankings that we interleave and expose to simulated users. A main difference among the interleaving methods is how they deal with the intersection between the compared lists – credit on items outside the intersection is assigned in the same way by all methods. To focus on that, we simulate pairs of rankings that contain the same set of items and only differ in the order. This is not uncommon in real online experiments where the compared treatments are alternative reorderings of the control system. To simplify, we make all the items in our simulations equal in user engagement potential, except one – let us refer to it as 'x' – with a higher action probability (e.g. because it is more relevant).

To create both an overlapping imbalance and a quality difference between rankers, we simulate a ranker A that places the high-engagement x at a random position in the bottom half of the ranking, and a ranker B that ranks x at a random position in the top half. B is

[1] Source code is available at https://github.com/amzn/debiased-balanced-interleaving.

better than A under any perspective, since it ranks the single more valuable item higher. On the other hand this configuration produces an overlapping imbalance favoring A: all the items between the repective position of x in B and A are one place lower in B than in A.

6.1.2 User behavior. User behavior consists of a browsing model and an engagement model, similar to simulations in prior work [7]. For item examination we take a cascade model [4], where examination probability decreases over ranking depth with nDCG decay: the probability that the user stops browsing at position k is $1-\log(k+1)/\log(k+2)$ [2]. For engagement, we consider two models:

a) An aimless user who engages with any examined item with equal probability. We use this model to check unbiasedness: an unbiased interleaved comparison should declare a tie for any two rankers when exposed to this random behavior model.
b) A purposeful user who takes action based on the engagement probability of seen items. We use this model to test sensitivity: a sensitive interleaving method should soon identify a statistically significant difference when exposed to this behavior, if high-engagement items are ranked at different positions.

6.1.3 Sessions. We simulate 100 sessions including 100 (imaginary) queries each – we intentionally take a small sample size to observe some variance; the number of queries per session is relatively high but is representative of the session notion in our use cases, and helps us get clean visualizations. For each query, the compared rankers A and B produce a ranking. Then for each session, we compute the credit of A and B by each interleaving method and take the average over sessions (and over queries to scale down results for readability). For our method, we count the attribution opportunity down to the last user examined item in the interleaved rankings (in our online experiments in section 7 we similarly count opportunity only for impressed items). We take a ranking size of 50 items, an engagement probability of 0.5 for all items except probability 1 for x and, to simplify, we count all user actions as producing unit credit of 1.

6.2 Unbiasedness

To evaluate the potential bias of different interleaving variants, we first examine their outcomes under random user behavior, as described earlier in the simulation setup. The top-left graph of Figure 2 shows the average credit under this regime: each point in the scatterplot represents a session, with credit of A in the x axis and credit of B in the y axis. A/B tests have half the number of points – randomly paired sessions – as traffic is evenly split between A and B. With random user engagement, an unbiased test should declare a tie for any two rankings in expectation: the dots in the graph should lie close to the $y = x$ line. We see this is definitely not the case for uncorrected balanced interleaving, evidencing the impact of the overlapping imbalance: ranker A is unfairly favored by a substantial extent and declared a winner in all sessions, because several items are one position higher in A than in B.

All other methods look aligned around $y = x$ with some variance, as far as the bare eye can catch. The bottom-left graph in Figure 2, however, uncovers a different story. Similar to prior work [6,19,24], the curves indicate in the y axis the number of times (out of 1,000 trials) an experiment finds a statistically significant difference (p-value < 0.05) between A and B, for different amounts of traffic in the x axis. The faster a curve grows, the less data it takes to draw significant conclusions, hence the higher the statistical power and

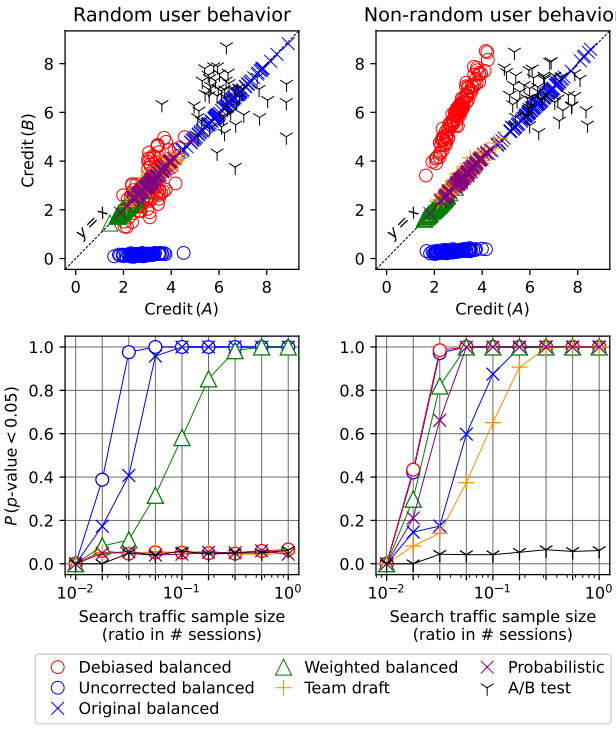

Figure 2: Bias and sensitivity analysis. To test unbiasedness, the left graphs show results under random user behavior. To test sensitivity, the right graphs show results under purposeful user behavior. The top plots display results for 100 sessions × 100 queries: each point represents one session and shows the average credit of ranker A (x axis) and B (y axis). The bottom graphs compare the statistical power of different methods: the simulation is repeated 1,000 times for different subsample sizes (number of sessions as a fraction of 100 in the x axis). For each sample size, the y axis shows how many times (as a fraction of 1,000 repetitions) the comparison between A and B is statistically significant (two-sided $p < 0.05$).

sensitivity of the method. Declaring a difference in this case is undesirable and reveals a bias – an unbiased comparison should have a flat power curve for random users. We see in the figure that our method is indeed unbiased and declares a tie for any sample size.

The figure also confirms that team draft and probabilistic interleaving do not suffer from the overlapping imbalance, as proved in the literature [3]. A/B tests are also unbiased as one should expect. On the other hand, the uncorrected balanced interleaving finds significant differences very early, confirming the strong drift in favor of A that is very visible in the top-left scatterplot. The other two variants of balanced interleaving show a clear bias too, even if it is not easy to see in the top-left scatterplot. For the original balanced attribution policy, this is a known fact in the literature [3,7,21]: it favors A as does the uncorrected variant. For the weighted balanced variant, the figure confirms that it does not properly fix the overlapping imbalance, even though it shows a weaker bias (the green curve grows more slowly than the two blue ones). Barely perceptible to the eye, this method favors B – a false positive, since customer behavior was random.

6.3 Sensitivity

We now examine sensitivity by considering the purposeful user model. The results are shown on the right side of Figure 2. In the top-right scatterplot we see that now, with informative user behavior, our proposed method correctly and unambiguously finds that ranker B is better (since B ranks the valuable item higher than A). Meanwhile the uncorrected balanced variant stubbornly keeps declaring A a winner – the overlapping imbalance against B is as strong as that. In contrast all the other methods seem to stay closer to the tie line $y = x$. Our method appears to be the most sensitive: B is distinctively declared a winner in *all* sessions.

The power curves (bottom-right) provide a more accurate picture of the sensitivity performance. Even if the difference is not apparent in the scatterplot, all methods find some significant difference in favor of the right winner B, though not as clearly as our method. Team draft is the weakest interleaved method in terms of sensitivity. The A/B test is even less sensitive, remaining flat in power. The compared rankings A and B are not that different after all – they contain the same items and only one differs in position, and the A/B test would need orders of magnitude more search traffic to detect this small difference, confirming that interleaving can achieve great improvements in data efficiency compared to traditional A/B testing. The uncorrected balanced variant seems to be as sensitive as our method, with one "small" caveat: it declares the wrong winner A with a misleadingly high confidence.

On the other hand, probabilistic interleaving overcomes the lack of sensitivity of team draft, thus meeting its purpose. For a closer comparison between our method and probabilistic interleaving, we test them on a known team draft edge case [3,7]: we simply take two rankings $A = \langle a, b, x \rangle$, $B = \langle b, x, a \rangle$ where, as in the previous simulations, the conversion rates are 1 for x and 0.5 for all other items, and we simulate the same amount of traffic. B is objectively better because it ranks the best item x higher than A. Figure 3 shows that all methods except (as expected) team draft detect the difference. We see however an advantage in our method achieving higher sensitivity than probabilistic interleaving, finding the difference with high certainty $P(p < 0.05) = 1$ at ~5 times less data (in the log-scale abscissa). While the probabilistic method relies on ranking randomization to detect the advantage of B over A, our method assigns to B all the credit of engagement with x. Therefore our method solves the team draft weakness in a more data-efficient and computationally inexpensive way than probabilistic interleaving.

6.4 Bias Analysis upon Alternative Policies

Our correction approach (dividing credit by credit opportunity) could also be applied to other similarly biased attribution policies in balanced interleaving, such as the original policy and the position-based weighting policy [3]. These do not enjoy the theoretical guarantees of our chosen variant, as credit attribution is not a proper distribution over two rankers – in particular, these policies attribute credit to both rankers for some items. We nonetheless test the correction on the alternative policies as additional empirical validation of our policy choice.

We see in Figure 4 (a) that the original policy is still biased by the overlapping imbalance: the power curve is far from flat under random user engagement. Weighted attribution appears as a potential unbiased alternative, slightly less sensitive (Figure 4 (b) right) than our chosen variant. This policy does not display any empirical advantage though, while having a difficult interpretation of what opportunity correction does (even on toy examples). This further confirms our preferred policy as an effective, simple and theoretically sound option for bias correction.

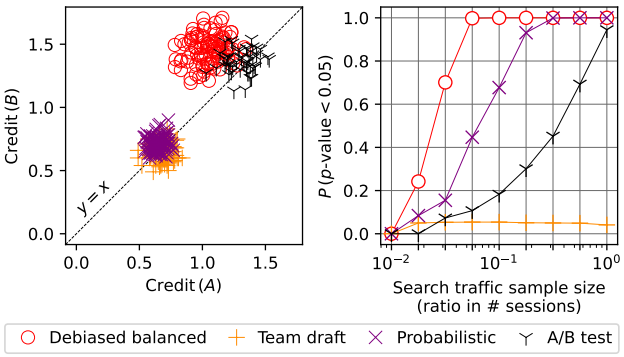

Figure 3: Sensitivity comparison on the team draft edge case under purposeful user behavior. The graphs display similar views as the right column of Figure 2.

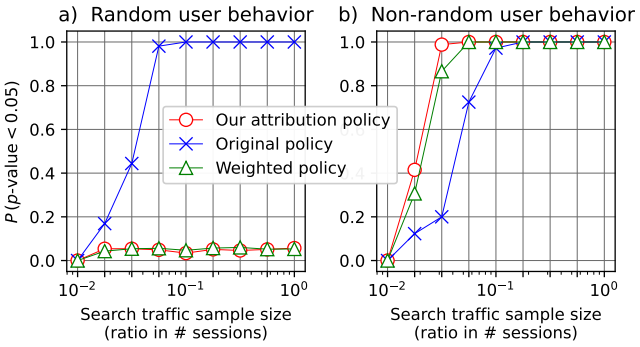

Figure 4: Applying the correction scheme on alternative credit attribution policies. The curve labeled "our attribution policy" corresponds to our proposed method, labeled "corrected balanced" in Figures 2 and 3.

6.5 Correlation with Business Metrics

Interleaved experiments are commonly handled as binary comparisons with three possible outcomes: win, lose or tie. Credit is not binary, but through elaborate attribution rules, the numeric credit value is difficult to interpret as a business metric. We empirically examine the correlation between the interleaved credit difference and the difference in number of actions (as a business metric proxy) in separate A/B tests with the same rankers.

For this purpose, we extend the simulation setup with different ranges of low and high engagement probabilities (other than 0.5 and 1) in the synthetic rankings, inducing a wider range of differences between A and B. Specifically, we range low engagement between 0.1 and 0.9 (by steps of 0.1), and high engagement from the low probability to 1, thus resulting in a total of 55 experiments. For each pair of rankers, we raise the number of sessions to 10,000 to make the correlations more perceivable. To make the observation symmetric we flip a coin to decide which is A and B in each experiment.

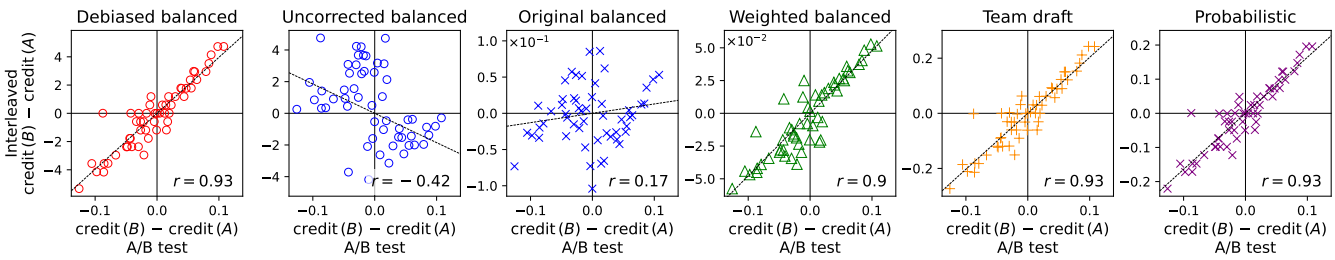

Figure 5: Comparing the numeric outcome of interleaved methods (difference in attributed credit, y axis) and separate A/B tests (difference in number of user actions, x axis) under non-uniform user engagement. Each point in the graphs corresponds to an "experiment" where a different pair of synthetic rankers are compared over 10,000 sessions × 100 queries. Pearson correlation and the zero-intercept trend line are shown for each plot.

Experiment #	# Sessions	Duration	Imbalance $\omega(B\|A)$
1	17.1 million	14 days	0.499
2	4.1 million	14 days	0.498
3	10.7 million	14 days	0.499
4	2.6 million	28 days	0.362
5	17.2 million	21 days	0.469
6	11.8 million	14 days	0.488
7	54.2 million	28 days	0.510
8	7.0 million	28 days	0.501
9	29.3 million	14 days	0.446
10	13.8 million	14 days	0.512

Table 1: Basic details of the online experiments. The highest incidence of overlapping imbalance is highlighted in red.

As shown in Figure 5, our proposed method, weighted balanced, team draft, and probabilistic interleaving all display a strong linear correlation with A/B tests. This suggests that the numeric outcome from interleaving can be indeed informative of the magnitude an A/B measurement would return. In contrast, the uncorrected and the original balanced variants do not display any meaningful correlation. In fact, the uncorrected variant has most points in the wrong quadrants due to the existing overlapping imbalance in our simulations, indicating a contradiction with A/B conclusions.

Also worth mentioning, we observe a considerable point by point similarity between team draft, probabilistic interleaving and our approach. This hints a fair degree of practical equivalence between these methods when comparing rankers that are different enough. In this context, our proposed method still stands as an advantageous choice with implementation and computation considerations.

7 ONLINE EXPERIMENTS

We assess our method with real Amazon search traffic, to evaluate fidelity in terms of agreement with A/B tests [8,24], and sensitivity in terms of statistical power. Based on our theoretical and simulation analysis in previous sections, team draft and probabilistic interleaving were discarded from implementation in the real life service. Limitations in sensitivity, implementation complexity, and/or runtime latency advised against these interleaving methods and their exposure to real customers. Our implementation does support a post-hoc comparison of our method and the original balanced interleaving, as well as traditional A/B testing.

The compared rankers in our experiments are repurposed from ten traditional A/B tests that have recently been run at Amazon search. The previous A/B experiments tested improved features for high quality product brands and fast shipping, as well as improved

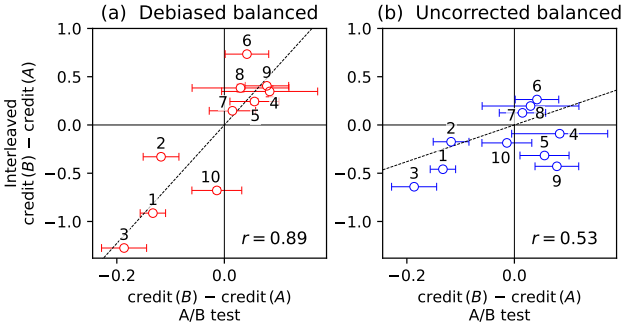

Figure 6: Fidelity of debiased and uncorrected balanced interleaving in online experiments. Similarly to Figure 5, we check the agreement of interleaved comparisons (y axis) with A/B tests (x axis), in terms of the credit difference in revenue between the compared rankers. The points are labeled with the corresponding experiments in Table 1.

modeling and data preparation methods. Table 1 has the basic details of these experiments, which are the same for interleaved and A/B tests. We report here the results using revenue as the performance metric [26]; we observe similar performance with other metrics such as clickthrough and conversion rates.

We define $\omega(B|A)$ to indicate the extent of overlapping imbalance for a experiment, which computes the ratio of overall attribution opportunities between rankers A and B:

$$\omega(B|A) = \frac{\sum_I \sum_k \mathbb{1}(\alpha_k(I) = B)}{\sum_I \sum_k \mathbb{1}(\alpha_k(I) = B) + \sum_I \sum_k \mathbb{1}(\alpha_k(I) = A)}$$

The more this ratio deviates from 0.5, the more of the impact by overlapping imbalance. As shown in Table 1, this ratio is generally not far from 0.5 in our experiments. However, those experiments that do exhibit significant overlapping imbalance (experiments 4, 5, and 9) are precisely those for which original balanced interleaving fails to agree with A/B tests, as we see next (Figure 6 and 7).

7.1 Fidelity

Figure 6 shows the experiment results. Like earlier in Figure 5 (top), each point shows two online evaluation results, interleaved and non-interleaved, for a pair of rankers. The coordinates reflect the credit difference in revenue between the two compared rankers in the corresponding experiment: x axis for the traditional A/B test, and y axis for the interleaved comparison. The right panel (b) shows the uncorrected balanced interleaving, and the left panel (a) shows the debiased interleaving comparison.

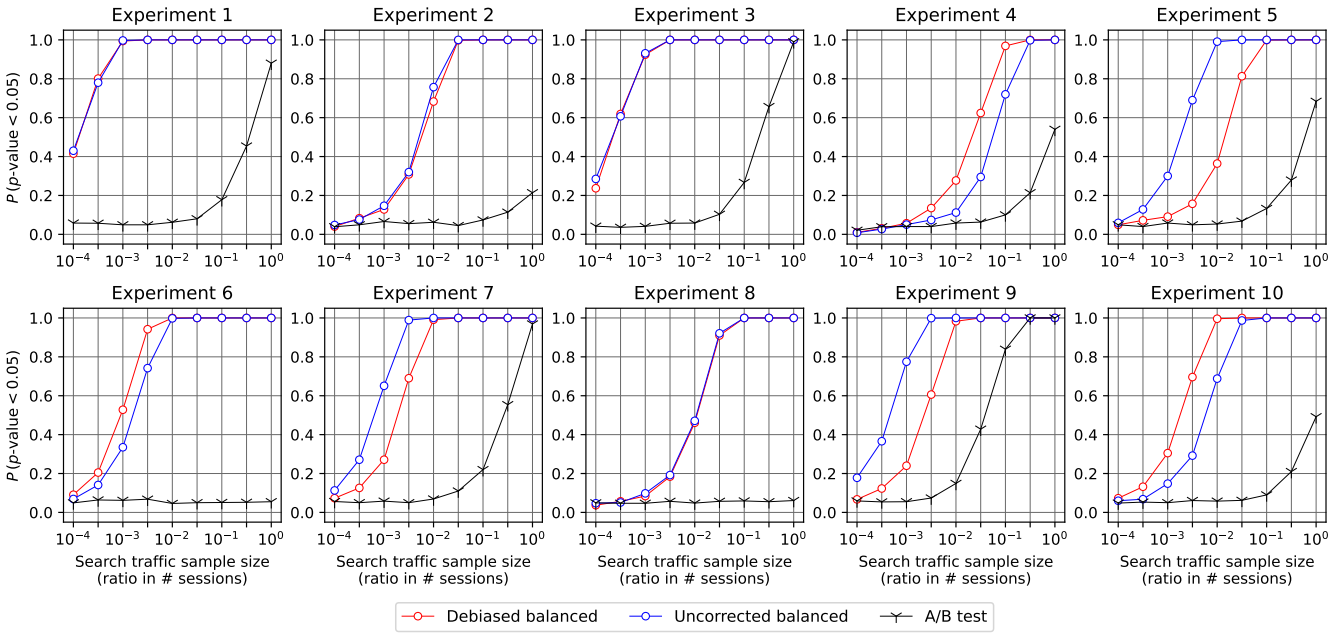

Figure 7: Sensitivity of interleaving and A/B tests in online experiments. As in Figures 2 and 4, statistical power is estimated by randomly subsampling experiment sessions, and counting the number of statistically significant results at $p < 0.05$ (y axis) for different subsample sizes (x axis). The faster the curve grows, the higher the statistical power.

We see that our debiased interleaving comparison agrees in sign with the A/B test in all ten experiments: the points in Figure 6 (a) lie in the top-right and bottom-left quadrants, meaning that the interleaved comparison declares the same winner as the A/B test in all cases. Furthermore, the interleaved results display a high linear correlation with the A/B measurements, which is potentially useful as a quantitative estimate of the business metric. In contrast, the uncorrected method in (b) has three points (experiments 4, 5, 9) in the "wrong" quadrant (right-bottom), where the interleaved comparisons disagree with A/B tests.

These results are obviously small-scale in the number of experiments (ten data points); but not so in traffic volume, which is in millions of sessions and hundreds of millions of requests per experiment. A more comprehensive statistical comparison would require thousands of experiments over hundreds of weeks, much beyond the scope of this paper. We nonetheless take these results as a positive proof of concept of the proposed method, and a good sign confirming that the proposed method works as expected.

7.2 Sensitivity

Figure 7 shows the statistical power of different evaluation methods. As in the simulations in Section 6, we estimate power by subsampling the experiment data at different proportions, and for each proportion we check the ratio of samples that reach a statistically significant conclusion. We see that our method does not incur any loss of power: both interleaving variants have equal or similar sensitivity overall. Though uncorrected measurements appear as more powerful in experiments 5 and 9, they actually return the wrong conclusions (bottom-right quadrant in Figure 6 (b): it is easy to achieve statistical significance with incorrect results. On the other hand, our method is much more data-efficient than A/B tests, reaching significant conclusions much more quickly. Our debiased interleaving is on average 60 times more powerful than A/B tests based on this set of experiments: $P(p\text{-value} < 0.05) = 1$ is reached ~ 60 times earlier in the x axis.

8 CONCLUSIONS AND FUTURE WORK

We propose a novel interleaving method that provides unbiased comparisons while sticking to the simplicity of balanced interleaving policy. Our method corrects the bias of balanced interleaving with theoretical guarantees, and still inherits its merits of sensitivity, user experience preservation, and inexpensive implementation. We empirically confirm the correctness of our method with synthetic and real search traffic, where the method consistently fixes the errors of balanced interleaving, while retaining the superior data-efficiency of interleaved comparisons with respect to traditional A/B tests. Our method stands the comparison against alternative unbiased methods, coming out as equally good or better in fidelity and sensitivity. In the comparison, the savings in implementation and computation complexity are a decisive advantage for large-scale experiments.

Our method has been implemented at Amazon search and tested at the scale of billions of search queries. For 10 large-scale online experiments at Amazon, our method reaches the correct conclusions in all of them, with an estimated 60x sensitivity gain over A/B tests. We furthermore observe a high correlation between the numeric interleaving results and corresponding A/B measurements.

For future work, we are interested in evaluating the stochastic transitivity for our method that defines the ideal relationship among multiple interleaving pairs. This paves the road for online learning to rank which compares multiple rankers (sequentially in multiple rounds) based on efficient pairwise comparisons [3,21]. Additionally, we can further improve the robustness of our method with a finer handling of position bias [27] in the estimation of attribution propensity, where we may over-penalize against the overlapping imbalance with credit opportunities not actually granted by users.

REFERENCES

[1] 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

[2] Ben Carterette. 2011. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR 2011)*. ACM, New York, NY, USA, 903–912. https://doi.org/10.1145/2009916.2010037

[3] 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

[4] Olivier Chapelle and Ya Zhang. 2009. A Dynamic Bayesian Network Click Model for Web Search Ranking. In *Proceedings of the 18th International Conference on World Wide Web* (Madrid, Spain) *(WWW 2009)*. ACM, New York, NY, USA, 1–10. https://doi.org/10.1145/1526709.1526711

[5] Jing He, Chengxiang Zhai, and Xiaoming Li. 2009. Evaluation of Methods for Relative Comparison of Retrieval Systems Based on Clickthroughs. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM 2009)*. ACM, New York, NY, USA, 2029–2032. https://doi.org/10.1145/1645953.1646293

[6] 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

[7] 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

[8] 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

[9] Kojiro Iizuka, Takeshi Yoneda, and Yoshifumi Seki. 2019. Greedy Optimized Multileaving for Personalization. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys 2019)*. ACM, New York, NY, USA, 413–417. https://doi.org/10.1145/3298689.3347008

[10] Thorsten Joachims. 2002. Optimizing Search Engines Using Clickthrough Data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Edmonton, Alberta, Canada) *(KDD 2002)*. ACM, New York, NY, USA, 133–142. https://doi.org/10.1145/775047.775067

[11] Thorsten Joachims. 2003. Evaluating Retrieval Performance using Clickthrough Data. In *Text Mining*, J. Franke, G. Nakhaeizadeh, and I. Renz (Eds.). Physica / Springer Verlag, 79–96.

[12] Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. 2013. Using Historical Click Data to Increase Interleaving Sensitivity. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (San Francisco, California, USA) *(CIKM 2013)*. ACM, New York, NY, USA, 679–688. https://doi.org/10.1145/2505515.2505687

[13] 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

[14] 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

[15] 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

[16] Donald R. J. Laming. 1986. *Sensory Analysis*. Academic Press, Cambridge, MA, USA.

[17] Harrie Oosterhuis and Maarten de Rijke. 2017. Sensitive and Scalable Online Evaluation with Theoretical Guarantees. In *Proceedings of the 26th ACM Conference on Information and Knowledge Management* (Singapore) *(CIKM 2017)*. ACM, New York, NY, USA, 77–86. https://doi.org/10.1145/3132847.3132895

[18] Harrie Oosterhuis, Anne Schuth, and Maarten de Rijke. 2016. Probabilistic Multileave Gradient Descent. In *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings (Lecture Notes in Computer Science, Vol. 9626)*, Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello (Eds.). Springer, 661–668. https://doi.org/10.1007/978-3-319-30671-1_50

[19] 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

[20] 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

[21] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2008. How Does Clickthrough Data Reflect Retrieval Quality?. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (Napa Valley, California, USA) *(CIKM 2008)*. ACM, New York, NY, USA, 43–52. https://doi.org/10.1145/1458082.1458092

[22] 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. https://doi.org/10.2307/2335942

[23] Anne Schuth, Robert-Jan Bruintjes, Fritjof Buüttner, Joost van Doorn, Carla Groenland, Harrie Oosterhuis, Cong-Nguyen Tran, Bas Veeling, Jos van der Velde, Roger Wechsler, David Woudenberg, and Maarten de Rijke. 2015. Probabilistic Multileave for Online Retrieval Evaluation. 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, 955–958. https://doi.org/10.1145/2766462.2767838

[24] 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

[25] 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

[26] Daria Sorokina and Erick Cantu-Paz. 2016. Amazon Search: The Joy of Ranking Products. 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, 459–460. https://doi.org/10.1145/2911451.2926725

[27] Yisong Yue, Yue Gao, Oliver Chapelle, Ya Zhang, and Thorsten Joachims. 2010. Learning More Powerful Test Statistics for Click-Based Retrieval Evaluation. 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, 507–514. https://doi.org/10.1145/1835449.1835534