A Counterfactual Framework for Seller-Side A/B Testing on Marketplaces

Viet Ha-Thuc
vhathuc@fb.com
Facebook Inc.
Menlo Park, CA

Avishek Dutta
avishek1013@fb.com
Facebook Inc.
Menlo Park, CA

Ren Mao
neroam@fb.com
Facebook Inc.
Menlo Park, CA

Matthew Wood
woodmd@fb.com
Facebook Inc.
Menlo Park, CA

Yunli Liu
yunliliu@fb.com
Facebook Inc.
Menlo Park, CA

ABSTRACT

Many consumer products are two-sided marketplaces, ranging from commerce products that connect buyers and sellers, such as Amazon, Alibaba, and Facebook Marketplace, to sharing-economy products that connect passengers to drivers or guests to hosts, like Uber and Airbnb. The search and recommender systems behind these products are typically optimized for objectives like click-through, purchase, or booking rates, which are mostly tied to the consumer side of the marketplace (namely buyers, passengers, or guests). For the long-term growth of these products, it is also crucial to consider the value to the providers (sellers, drivers, or hosts). However, optimizing ranking for such objectives is uncommon because it is challenging to measure the causal effect of ranking changes on providers. For instance, if we run a standard seller-side A/B test on Facebook Marketplace that exposes a small percentage of sellers, what we observe in the test would be significantly different from when the treatment is launched to all sellers. To overcome this challenge, we propose a counterfactual framework for seller-side A/B testing. The key idea is that items in the treatment group are ranked the same regardless of experiment exposure rate. Similarly, the items in the control are ranked where they would be if the status quo is applied to all sellers. Theoretically, we show that the framework satisfies the stable unit treatment value assumption since the experience that sellers receive is only affected by their own treatment and independent of the treatment of other sellers. Empirically, both seller-side and buyer-side online A/B tests are conducted on Facebook Marketplace to verify the framework.

CCS CONCEPTS

• **Mathematics of computing** → *Hypothesis testing and confidence interval computation*; • **Information systems** → *Evaluation of retrieval results.*

KEYWORDS

A/B testing; counterfactual; ranking evaluation; marketplace; design of experiments

ACM Reference Format:
Viet Ha-Thuc, Avishek Dutta, Ren Mao, Matthew Wood, and Yunli Liu. 2020. A Counterfactual Framework for Seller-Side A/B Testing on Marketplaces. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3397271.3401434

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07.
https://doi.org/10.1145/3397271.3401434

1 INTRODUCTION

Many consumer products on the Internet are two-sided marketplaces. For instance, Amazon, Alibaba, and Facebook Marketplace connect buyers and sellers. LinkedIn and Facebook newsfeeds are products where content consumers and producers interact. Sharing economy products like Uber, DiDi and Airbnb make connections between passengers and drivers or guests and hosts. The search and recommender systems behind these products are typically optimized for consumer-oriented metrics like purchase rate, click-through rate, booking rate, etc. [4, 12] (For the rest of the paper, the term *buyer* is used in a broad sense indicating all types of consumers, such as buyers on Amazon/Alibaba/Facebook Marketplace, content consumers on newsfeeds, or passengers/guests on sharing economy products. The term *seller* broadly refers to the producer side of the marketplace, like sellers, content producers, drivers, or hosts.)

However, it's also crucial to optimize for seller-side objectives. Some examples of seller-side objectives include the percentage of sellers having success when using the product (e.g., having at least one item sold, one home reservation, or one ride) and seller retention. Ensuring that sellers have a great experience and increasing the likelihood they will return are essential to the long-term success of the products. To the best of our knowledge, there has been little work focusing on these objectives. One of the reasons is that unlike the buyer side, measuring the impact of a ranking change on sellers is challenging.

To demonstrate the challenge of measuring the impact of a ranking change on sellers, let us consider the Facebook Marketplace product where buyers can discover items for sale from local communities or businesses. When a buyer visits Marketplace on Facebook,

a marketplace feed shows a ranking of items by personalized relevance (the left screenshot on Figure 1). Assume we run a simple seller-side A/B test evaluating the causal effect of a ranking change on *sellers*. As a concrete example, the ranking change boosts the ranking positions of items from first-time sellers on Marketplace.

In this A/B test, the treatment group contains a 1% sample of randomly-selected sellers, and the items from the new sellers in this group are given a boost to their ranking scores. The control group contains another 1% sample of randomly-selected sellers, and their items do not get a boost. Because of the boost, the items from the sellers in the treatment group (in short, the items in the treatment) are ranked higher in the feed. Thus, the new sellers in the treatment group get a better experience compared to the ones in the control group. However, when the boosting is ramped up to 100% (i.e., to all new sellers), the items in the original treatment group will be ranked lower because a larger fraction of items are getting boosted now. Thus, the impact on the sellers when experimented at 1% is artificially inflated, and the seller A/B test result is incorrect.

To resolve the cannibalization effect above, we propose a novel counterfactual framework for seller-side A/B testing. It is based on a counterfactual property: when experimented at a small percentage of sellers, the items in the treatment group are ranked at where they would be if the treatment is ramped to 100% of the sellers. Similarly, the items in the control are ranked where they would be if the status quo is applied to all sellers. Thus, the difference between treatment and control is independent of what applies to the rest of the sellers.

To verify the counterfactual framework, we run a seller-side A/B test on the framework in which 1% of the sellers get the new-seller boosting treatment mentioned above and 1% of control sellers do not get the boosting treatment. In the middle of the experiment period, the boosting treatment is launched to the remaining 98% of sellers. The experimental results show that seller-side metrics of the treatment and control groups before and right after the launch are consistent. This confirms the cannibalization effect does not happen. Furthermore, we A/B test the ranking change on both the buyer and seller sides separately. We observe that for metrics measurable from both sides (e.g., the number of buyer-seller interactions), the experimental results from the buyer-side and the seller-side A/B tests are also consistent. As of writing this paper, the counterfactual framework has been used for on-going seller-side experiments on Marketplace and other Facebook products.

The rest of the paper is organized as the following. Section 2 gives background information on the Facebook Marketplace product, which is the case study used in this paper. Section 3 explains in detail why it is challenging to measure the causal effect of ranking on sellers. To solve this challenge, a counterfactual framework for seller-side A/B testing is proposed in Section 4. Then in Section 5 we present empirical results validating the counterfactual framework. Previous work in the literature related to ours is reviewed in Section 6. Finally, in Section 7 we give our concluding remarks.

2 BACKGROUND

2.1 Product Overview

This section gives an overview of Facebook Marketplace, the case study and product on which all experiments are conducted in this paper. Facebook Marketplace is a social shopping platform where users connect to buy and sell various items from electronics to cars to apparel. Facebook Marketplace is available in many countries around the world and there are hundreds of millions of people using Marketplace on Facebook.

From the home screen of the Facebook app, a person can visit Marketplace by tapping on the shop icon at the bottom of the app. Then the person is presented with the Marketplace browse feed, a ranked list of the items in the person's local area as well as shipped items (Figure 1 left screenshot). The items are sorted by relevance to the person's interests. When the person finds a relevant item, clicking on the item opens an item detail page where more information about the item like description and price is shown (center screenshot). If the person finds the item interesting, they can message the seller from the item detail page to ask for further information, negotiate price, or arrange to buy the item (right screenshot).

Besides browsing the recommended items, people can also search for their items of interest by entering queries on the search box (Figure 2). Then a search result page appears with a ranked list of items. As on the browse feed, people can click on the items on the result page to land on the detail page and initiate a message to the seller.

One challenge in both the search and recommendation systems above is how to optimize the item ranking given a buyer, context (e.g., time, location, etc.), and query (in the case of search). We will discuss this in the next subsection.

2.2 Ranking Optimization

A common approach to ranking optimization in many industrial systems is to estimate the probability of some events of business interest, such as a buyer clicking on an item, a buyer liking or sharing an item to friends, or a buyer sending a message to the seller (buyer-seller interaction) [1, 4, 15]. Each of the probabilities is predicted by a machine learning model taking into account many features based on the item, buyer, context (i.e. location, time, etc.), and query (in the case of search ranking). These models are usually trained on historical actions in the log data. For instance, the training data of the click model includes past impressions with clicks as positive instances and impressions without clicks as negative ones. The individual component scores (the predicted probabilities) are then combined into a single score by a value model, which determines the relative importance of the events.

As shown above, rankings are typically optimized towards the value that buyers get from the items. However, there is another aspect in two-sided marketplaces: the values that *sellers* receive. Thus it is crucial to also optimize the rankings for seller-side objectives. An example of such objectives is the percentage of sellers that have success with the system. The success can be defined as receiving at least a certain number of messages, bookings, rides, etc. on a daily basis. Ideally, the system should maximize the number of successful sellers. Another example could be fairness across different groups of sellers [6], like new sellers versus existing ones.

Besides the seller-side objectives derived directly from the actions on the rankings, it is also essential to optimize the rankings for the future behavior of the sellers, such as retention and future

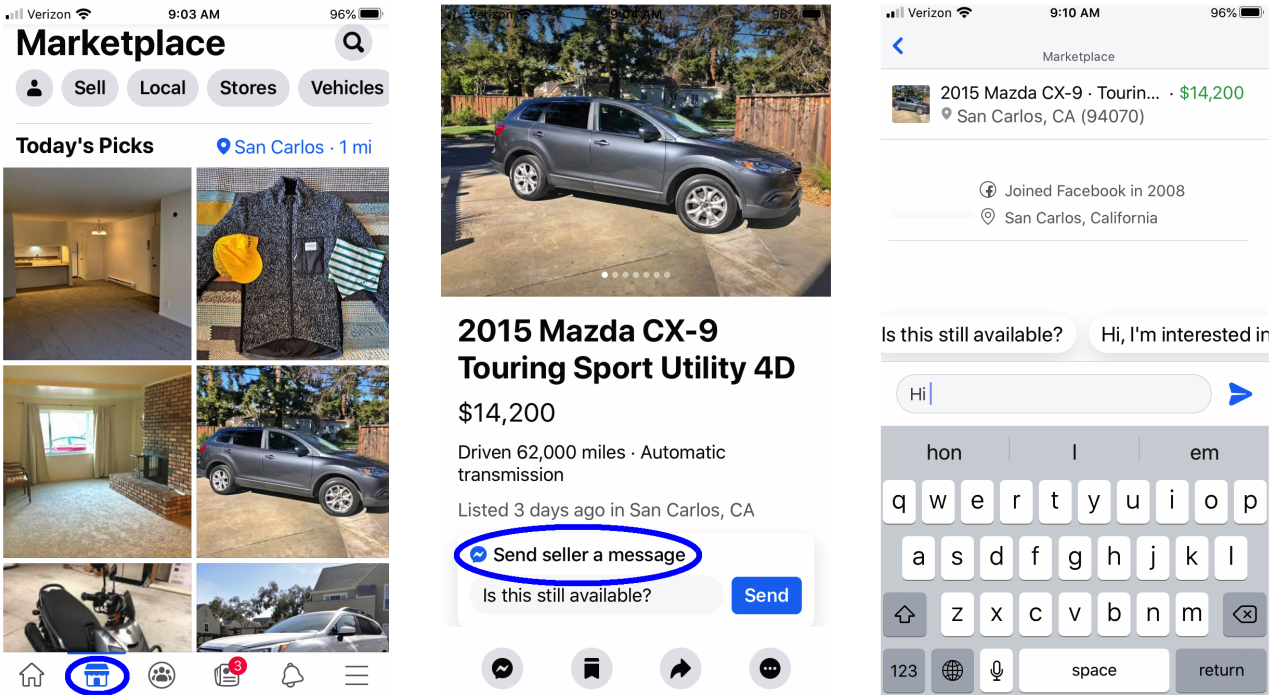

Figure 1: Left: a screenshot of Facebook Marketplace browse Feed. Center: an item detail page. Right: a buyer messages seller regarding to the item.

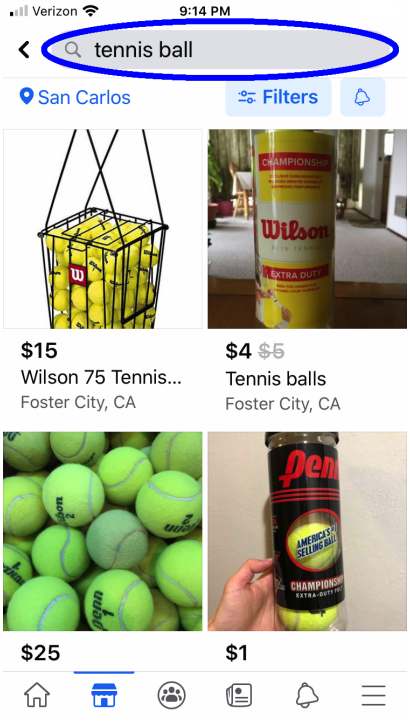

Figure 2: Facebook Marketplace Search

listings. Note that purely optimizing on buyer satisfaction, as mentioned above, likely leads to directing traffic to only a small portion of the total sellers. This causes the sellers (especially the new sellers) who do not get enough distribution to churn out. In the industrial setting, having more sellers return to produce more available items or services in the future is important to businesses [12]. The larger inventory will, in turn, benefit buyers in the future as well.

However, unlike the buyer-side objectives, one key challenge of optimizing seller-side objectives is measuring the causal impact of a ranking change on sellers. This will be discussed in detail in the next section.

3 CHALLENGE OF MEASURING IMPACT OF RANKING ON SELLERS

For almost all industrial search and recommender systems, before deploying a new ranking function, it is common practice to run A/B tests comparing the new version with the one currently in production [14]. In a typical A/B test, a random portion of the overall users (i.e., buyers in the case of Facebook Marketplace) is served with the new function (treatment). At the same time, another random portion of the buyers is served with the current production function (control). Usually, the sizes of the treatment and control groups are small (e.g., one or several percent) in order to minimize the opportunity cost and the negative impact on the user experience in the case that the new function is worse than the current one.

Comparing how the buyers in the two small groups engage with the product presumably indicates the difference between two

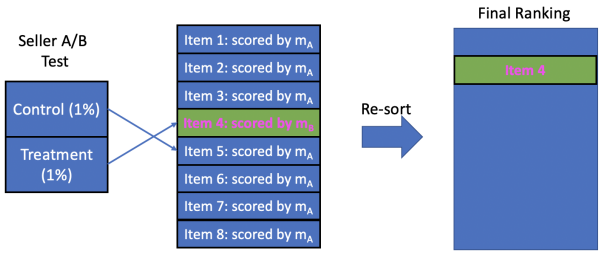

Figure 3: A simple seller-side A/B test where items whose sellers in the treatment group are ranked by a different ranking model.

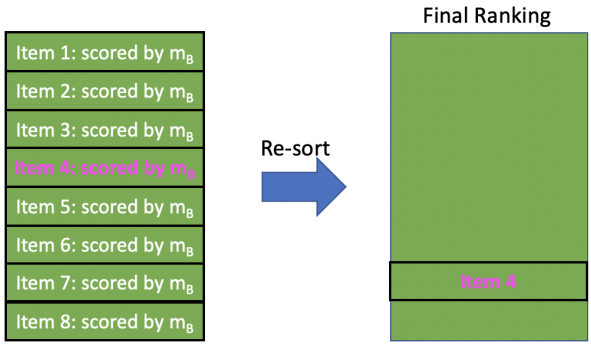

Figure 4: When the experimented ranking model is launched to all sellers, the items in the original seller-side A/B test (e.g., item 4) are ranked at different places. Thus, the seller-side effect when experimented at 1% is not the same as when launched to 100% sellers.

outcomes that we cannot observe simultaneously: the buyer engagement if the new ranking function is applied to everyone versus the buyer engagement if the status quo is maintained in the whole universe. The correctness of the A/B test (the equivalence between the observed and the counterfactual differences above) is based on a fundamental assumption, known as "stable unit treatment value assumption" in statistical inference theory [13, 18, 19]. The basic idea of the assumption is that the level of buyer engagement in either the treatment or control group is only dependent on the treatment these buyers received (the ranking function in this case) and independent of the ranking function applied to the buyers outside of the group.

While the stable unit treatment value assumption generally holds in the typical buyer-side A/B test mentioned above, it is definitely not the case in the seller-side A/B tests. To demonstrate this, let us consider ranking model m_A currently running in production that is solely optimized on buyers' objective and a new ranking model m_B, jointly optimized on both buyers' and sellers' objectives. We compare them on some *seller-side metrics*, e.g., the percentage of unique sellers getting messages from buyers and the seller retention rate. The most straightforward way to measure the impact of the new ranking model is to create a "naive" A/B test on the seller side. As demonstrated in Figure 3, for this test, items from the sellers in the treatment group are scored by m_B while the rest (including the items from the sellers in the control) are scored by m_A. Finally, the items are sorted by their scores into a single ranking. Then, we compare the seller-side metrics of the treatment and control groups.

After the experiment period, assume we launch the ranking model m_B from 1% to 100% (Figure 4). Even though the score of item 4 does not change (between the experiment period and after launching), the scores of the others in the set do change due to switching from using m_A to m_B. As a result, the position of item 4 will be different (even if all features stay the same). So, this naive test clearly violates the stable unit treatment value assumption mentioned above. Positions of the items in the treatment group (thus the experience that their sellers receive) do not only depend on the ranking function applied to them but also the ranking function applied to the other items.

To further illustrate the point, let's use a concrete example where m_B gives some boost to the items from new sellers who use the product for the first time. The rationale of the boosting is to avoid the cold start problem for these sellers [21]. This improves the experience of the new sellers, thus increasing the chance that they will stick with the product. When experimented at 1% of the sellers, items from the new sellers in the treatment group should be ranked high because of the boost.

However, when the boosting is ramped up 100% or in the back-test setting (where the boosting is ramped to 99% of the sellers and the status quo remains for just 1%), the same items in the test group of the original A/B test are ranked lower. This is because there are significantly more items getting boosted now (100 or 99 times more). Thus, the impact on the new sellers when experimented at 1% is not the same as when launched to 100% or back-tested. In other words, the effect in the original A/B test (also referred to as the pre-test) is artificially inflated, and the seller A/B test result is incorrect. In the next section, we propose a counterfactual experiment framework that can avoid the effect mentioned above.

4 A COUNTERFACTUAL FRAMEWORK FOR A/B TESTING

To compare two ranking models m_A and m_B on seller-side metrics, ideally for every request, we generate two rankings of the same set of items completely ranked by the models (Figure 5). Then, we hypothetically show both rankings to the user and compare them. This, of course, cannot happen in reality since we can only show one ranking to the user.

Inspired by this idea, however, we propose a counterfactual experiment framework for seller-side A/B testing. In this framework, we generate two complete rankings by the two models (See the upper part in Figure 6). To combine them into a single ranking, for the items in the control group (denoted as C_1 and C_2), we get their positions from the ranking by m_A. For the items in the treatment group (denoted as T_1 and T_2), however, we get their positions from the ranking by m_B, as demonstrated in the center box in the figure. In pre-tests, we enforce the positions of these items on the ranking by m_A to get the final ranking (the lower-left box). Similarly, in back-tests, we enforce the positions of the items on the ranking by m_B. Thus, in either case, the items in the treatment group would

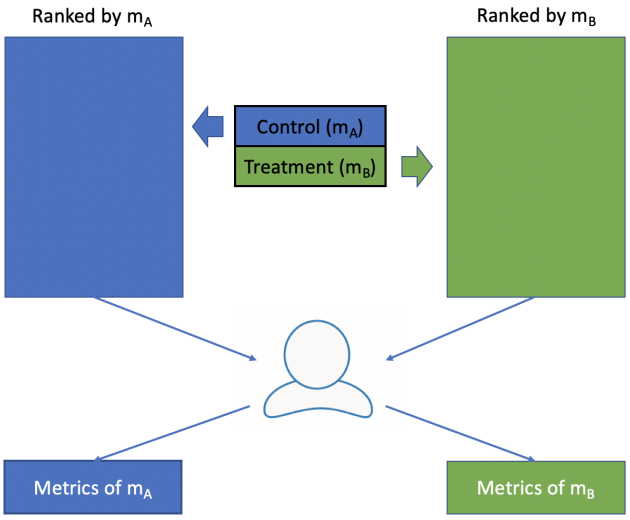

Figure 5: A hypothetical comparison between two models on seller-side metrics.

be ranked at the positions as if all results were ranked by m_B, regardless of which model is applied to the rest of the sellers. At the same time, the items in the control group would be ranked at the positions as if all results were ranked by m_A.

In the case of collision, e.g., ranking position of C_1 by m_A is the same as ranking position of T_1 by m_B, we can randomly move one to right below the other. Because the sizes of treatment and control groups are small (usually 1%-2%), collisions are rare. For example, if the treatment and control groups are both 1%, then the probability that there are items from both groups in top-10 is less than 0.914%, as shown in Equation 1. Furthermore, the chance that one item in the treatment group is ranked at the same place as another item in the control group is significantly smaller. Given the top-10 contains an item in the treatment and an item in the control and assuming the two rankings produced by m_A and m_B are uncorrelated, the collision only happens in 10% of those cases. In reality, m_A and m_B are usually highly similar. The more similar the two functions are, the smaller the likelihood that collisions happen (in the extreme case that they are the same, there is zero chance that two different items are ranked at the same position).

$$
\begin{aligned}
&P(\text{top-10 contains both treatment and control}) \\
&\approx P(\text{top-10 contains treatment}) * P(\text{top-10 contains control}) \\
&= (1 - \prod_{i=1}^{10} P(\text{item i not in treatment})) \\
&* (1 - \prod_{i=1}^{10} P(\text{item i not in control})) \\
&= (1 - 0.99^{10}) * (1 - 0.99^{10}) \\
&= 0.00914
\end{aligned} \tag{1}
$$

With this counterfactual framework, the comparisons between control and treatment in the pre-test (experiment mode) and the back-test (post-launching mode) are the same. In either test, the metrics on the sellers in the control group (e.g., percentage of unique sellers getting messages from buyers) are the same as in the case where the whole ranking by m_A is shown to the buyer since the sellers in the control are a random subset. Similarly, the metrics on the sellers in the treatment group are the same as when the whole ranking by m_B is shown to the buyer.

From a theoretical perspective, let us first adapt the traditional stable unit treatment value assumption for the seller-side A/B testing: the experience that sellers receive is only affected by their own treatment (ranking function in our specific case) and independent of the treatment of any other seller. It is evident that the counterfactual framework satisfies the assumption since the ranking position of each item is guaranteed to be the same regardless of what applies to other sellers in the whole universe.

The idea can be easily generalized to the case where we have multiple arms (i.e., test groups) in the experiment instead of just two arms. Assume we would like to compare the impacts of N ranking models $m_1, m_2 \ldots m_N$ on sellers. We generate N test groups in which each group contains randomly selected sellers from the overall universe. To limit the chance of collision mentioned above, we use a heuristic that the total size of all test groups is not bigger than 30% of the universe. At the ranking time, the process is similar to the one in Figure 6, except that the framework now generates N complete rankings (instead of two). For each item whose seller belongs to one of the test groups, its position is taken from the corresponding ranking. The final rankings of the pre-test and back-test are generated exactly the same way as in the figure. It is straightforward to verify that this generalized framework also satisfies the stable unit treatment value assumption mentioned earlier in this section.

5 EMPIRICAL RESULTS

This section shows various empirical evidence confirming the correctness of the counterfactual experiment framework.

5.1 Consistency Between Before and After Launching

To verify the framework, we run a seller-side A/B test evaluating the causal impact of a simple ranking change on Facebook Marketplace browse feed. The change aims to improve the experience of new sellers. Given a ranking by the first-stage ranker, among the items in the top 40 that are from new sellers not receiving any messages in the last 24 hours, we boost up to four items with the highest scores to positions 2 through 5 (0-indexed). We limit boosting to items in the top 40 to make sure the boosted items have reasonable quality, and we do not boost to the first two positions to alleviate the impact on buyers' experience. This simple re-ranking rule keeps the logic highly intuitive and allows us to easily control the trade-off between buyers' experience and sellers' experience.

The treatment group contains a random sample of 1% of sellers. In the pre-test, only the items from these sellers are impacted by the re-ranking rule above, but they are positioned as if the rule were applied to all items. The items from new sellers in the treatment

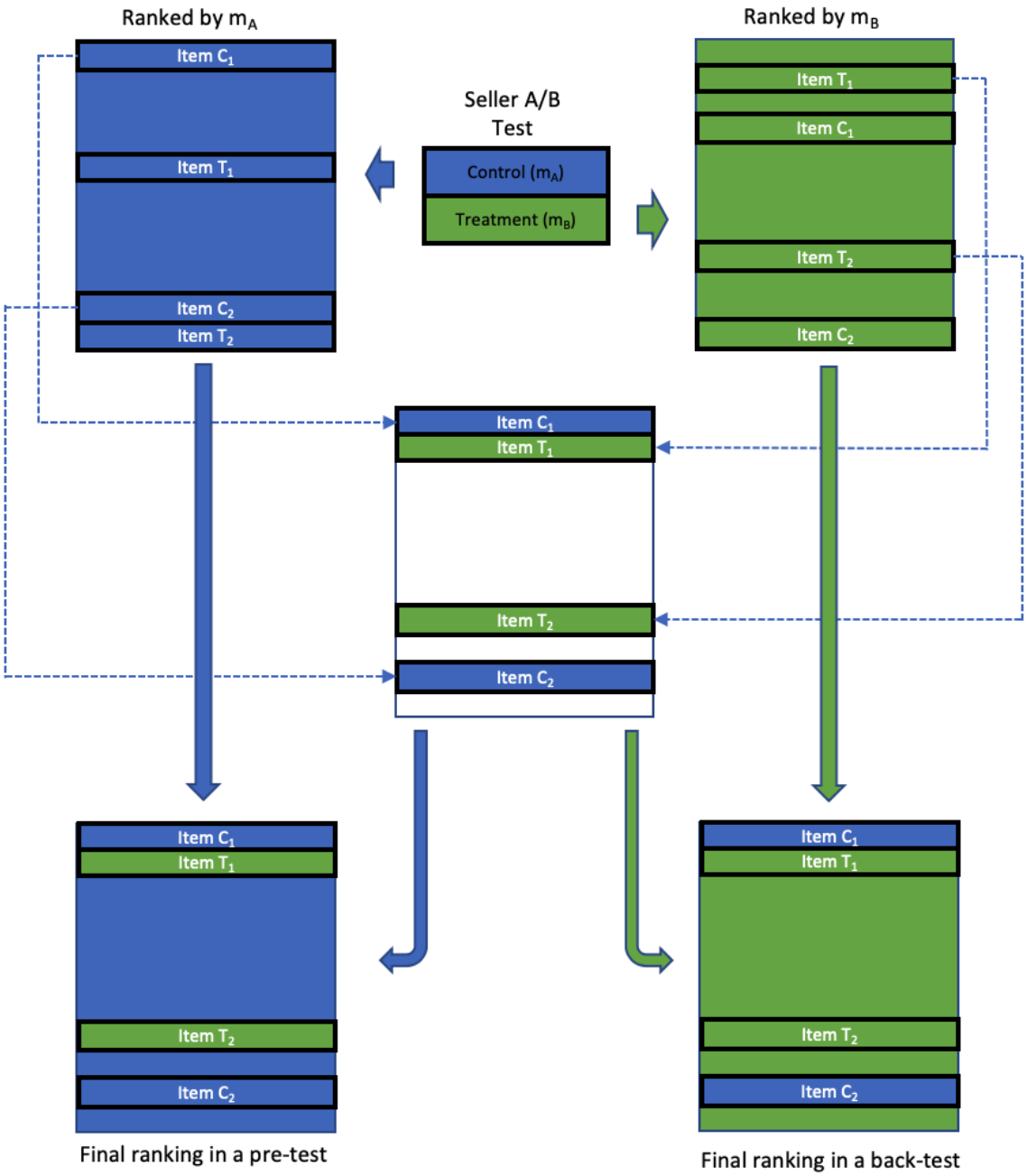

Figure 6: In the final rankings, positions of items from the sellers in the treatment group (T_1 and T_2) are the same as if all results were ranked by m_B and positions of items from the sellers in the control group (C_1 and C_2) are the same as if all results were ranked by m_A.

satisfying the conditions are up-ranked, and the other items in the treatment could potentially be ranked lower compared to the original ranking. The control group contains another 1% of sellers, and the re-ranking rule does not apply to their items (as well as the remaining 98%). We compare the two groups on two seller-side metrics: *messages received* and *seller success rate* (See Table 1). Since the ranking change improves the experience of the seller segment that are least likely to be successful, the seller success rate is expected to increase. The goal is to improve the metric while avoiding or minimizing the impact on the messages received.

In the middle of the experiment period, we launch the ranking change to the remaining 98% of sellers, i.e., we switch the pre-test into a back-test. Then, we observe the differences between the treatment and control groups before and after the launch. On the

Table 1: Seller-side metrics

Metric	Description
Messages received	The number of messages (buyer-seller interactions) that a seller receives on a given day.
Seller success rate	The percentage of unique sellers receiving at least one message on a given day.

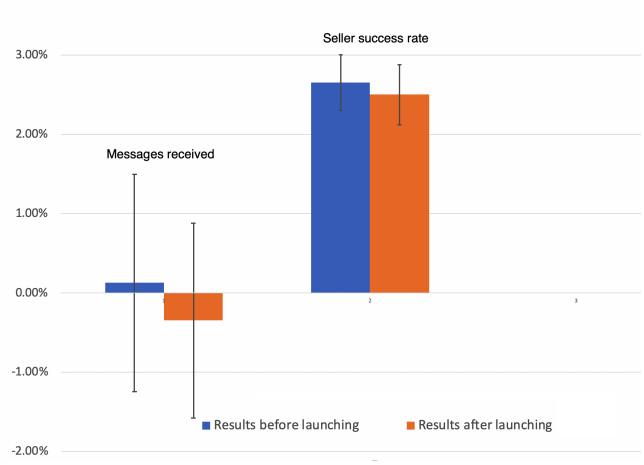

Figure 7: Consistency between before and after launch. The left bars indicate the relative differences between treatment and control groups in the number of messages sellers receive. The right bars are relative differences in the seller success rate.

left of Figure 7, the blue bar indicates the relative difference between the numbers of messages the sellers in the treatment group and the sellers in the control group received in the pre-test. The orange bar is the relative difference in the back-test. The black bars show the confidence intervals. As shown, the differences between the metrics before and after launch are not statistically significant.

The blue and orange bars on the right of the figure show the relative improvements of the treatment over the control in terms of seller success rate in the pre-test and back-test periods, respectively. Both pre-test and back-test show a statistically significant improvement in seller success rate of around 2.5%. So, on both seller-side metrics, the experimental results before and after launch are consistent.

5.2 Consistency Between the Buyer-Side and Seller-Side A/B Tests

The ranking change above can also be A/B tested on the buyer side. In the buyer-side A/B test, the rule is applied to the buyers in the treatment group and not applied to the ones in the control. Then, we can compare the two groups on the buyer-side metrics like the number of messages the buyers send. Note that if the stable unit treatment value assumption holds on both buyer and seller sides, the difference in the number of messages the buyers send

Table 2: Boosting strategies to improve the seller's experience. Treatment 1 is the least aggressive, and Treatment 3 is the most aggressive.

Boosting strategy	Description
Control	No boosting
Treatment 1 (40-4-2)	Boost up to 4 items in top-40 from new sellers to positions starting from 2.
Treatment 2 (100-4-2)	Boost up to 4 items in top-100 from new sellers to positions starting from 2.
Treatment 3 (100-6-0)	Boost up to 6 items in top-100 from new sellers to positions starting from 0.

(on the buyer-side test) should be consistent with the difference in the number of messages the sellers receive (on the seller-side test). Since it is straightforward to see that the assumption holds on this buyer-side A/B test, checking the consistency between the two tests verifies the correctness of the seller-side test.

Besides the re-ranking rule mentioned earlier (Treatment 1 in Table 2, we also experiment with some other variants (Treatments 2 and 3). Among the three boosting variants, Treatment 1 is the least aggressive, and Treatment 3 is the most aggressive. For each variant, we conduct A/B tests on both the buyer and seller sides. Then, we verify if the trend across the three boosting variants on the buyer-side is the same as the trend on the seller-side and if the buyer-side result is consistent with the seller-side result across each variant.

Figure 8 shows the differences between each treatment group over the control on the metric of messages sent/received. For each boosting strategy, the blue bar shows the difference measured by the buyer-side experiment, while the orange bar is the difference from the seller-side experiment. In Treatment 1, both of the experiments give neutral results. As expected, Treatment 2 shows significant regressions and Treatment 3 results in even larger (and consistent) regressions on both buyer-side and seller-side experiments. Moreover, across the boosting variants, the difference in magnitudes on the buyer-side and seller-side are well within their respective confidence intervals. The consistency between the buyer-side and seller-side experiments confirms that the seller-side experimental results are reliable.

6 RELATED WORK

Related to the challenge of A/B testing on the seller side is the network effect in network bucket (A/B) testing [5, 11]. The intuition behind this concept is that in social networks, a user's behavior is not only influenced by their own experience but also their friends'. As an illustrative example, suppose a newsfeed system currently only shows cat photos and no dog photos. Now, we run an A/B test on a small portion of newsfeed consumers in which the users in the treatment group only see dog photos on their feeds. Since most of their friends are not in the treatment group, these users only see photos with little engagement (e.g., likes, comments, etc.) from friends. Thus, they do not engage with the photos themselves. However, if the change is applied to all newsfeed consumers, everyone will only see dog photos, and their overall engagement will be the

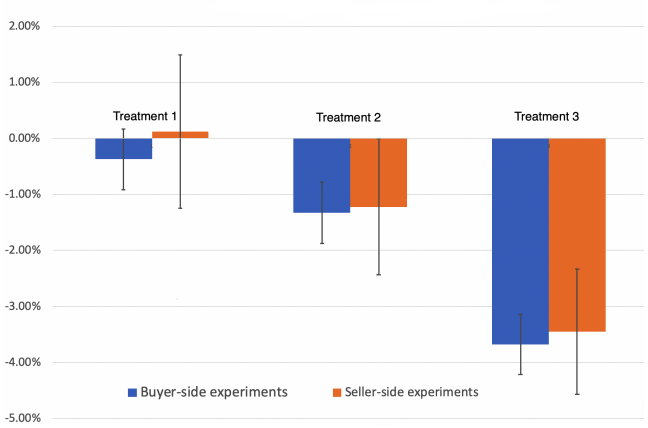

Figure 8: Consistency between buyer-side and seller-side experiments. The blue bars indicate the relative differences in the number of messages buyers send. The orange bars are the relative differences in the number of messages sellers receive.

same (assuming dogs and cats are intrinsically equally interesting). Thus, the stable unit treatment value assumption does not hold in the social network setting.

A typical solution to alleviate the network effect on network bucket testing is cluster-based randomization approach [2, 20, 24]. Before the online A/B test, users are clustered based on their connections. Then, the randomization is performed at the cluster level instead of the individual level, i.e., all users in the same cluster are either in the treatment or control group. With the clustering structure, there are relatively few connections across the clusters. Thus, the network effect is reduced.

Note that the spill-over effect above and the challenge of the seller-side ranking A/B tests described in Section 3 are similar but not the same. In the example above, the spill-over effect on newsfeed consumers in the treatment is indirect and via social influence. The experience of the remaining consumers has an impact on the photos, which in turn influences the engagement of the consumers in the treatment group on them. Usually, the changes have to be rather significant to spill over (100% dog photos versus 100% cat photos in the illustrative example). On the seller-side A/B test example, the ranking function used on the remaining 99% sellers has a direct impact on positions of the items in the treatment since all the items compete with each other on the same ranking. Thus, even if the ranking function m_B is just slightly different to m_A (in Figures 3 and 4), switching from m_A to m_B on the remaining will likely change positions of the item in the original treatment group.

Even given the difference, interestingly, the cluster-based randomization approach can still be adapted to solve the seller-side ranking A/B testing in a special case: buyers and sellers form a large set of communities, and the buy and sell activities only happen within each community. In this setting, during a seller-side A/B test, the new ranking function is only applied to sellers in a random set of the communities. Since a buyer only sees items from the sellers belonging to the same community, the ranking exposed to the buyer does not depend on the ranking function used in other communities. However, in many real-world marketplaces, to guarantee the communities are relatively isolated, the communities have to be large. For instance, using Uber as an example, an interaction could be reasonably expected between a passenger and a driver within a city or a region. Thus, the communities have to be at least at the level of cities. The large sizes of the communities lead to a small number of randomized units. Thus, this would significantly reduce the statistical power of the test and increase type II error rate [3]. In more open marketplaces like Alibaba, Amazon, Facebook Marketplace (where shipping across cities or even countries is available), and Airbnb (where a user in Asia might book a house in Europe), the cluster-based randomization approach is almost infeasible. The counterfactual framework, on the other hand, does not require any clustering structure among users. Indeed, it can work on the most open marketplace where any buyer can potentially connect to any seller. Moreover, on the framework, the randomization is performed on individual sellers, thus achieving high statistical power of the A/B tests.

Another research area in the literature also related to the work in this paper is Multi-Objective Ranking Optimization (MORO) [9, 22]. MORO is an approach to learning ranking models that optimize multiple objectives simultaneously, which is a common setting in many industrial search and recommendation systems. MORO has been a well-studied research area in the context of Web search [8–10, 22] and product search [7, 16] where relevance to the query, freshness, and user actions (like purchase) are all essential. Similarly, on recommender systems where users can have multiple possible actions such as clicking, liking, sharing (an item), applying and following (a job), MORO is also a natural approach [17, 23]. However, there has not been a lot of research focusing on jointly optimizing both buyer-side and seller-side objectives.

7 CONCLUSIONS

In this paper, we present a key challenge in an emerging direction of optimizing rankings on seller-side objectives. In many industrial systems ranging from commerce products (e.g., Amazon, Alibaba, and Facebook Marketplace) to newsfeeds (e.g., on Facebook and LinkedIn) to sharing economy (e.g., Airbnb, Uber, Ola, and DiDi), improving sellers' experience and retention is an important goal. However, it is challenging to A/B test the causal impact of ranking changes on sellers.

To overcome the challenge and open up the opportunity to introduce the seller-side objectives in the ranking optimization, we propose a novel counterfactual framework for seller-side A/B testing. It is based on a counterfactual property: when experimented at a small percentage of sellers, the items in the treatment group are ranked at where they would be if the treatment is ramped to 100% of the sellers. Similarly, the items in the control are ranked where they would be if the status quo is applied to all sellers. Thus, the seller-side metrics on the treatment or control are independent of what applies to the rest of the sellers.

Theoretically, we adapt the stable unit treatment value assumption traditionally used in standard buyer-side A/B testing for the seller side. We show that the proposed framework satisfies the

assumption. Empirically, our seller-side online A/B tests show consistent results between pre-tests (before launching) and back-tests (after launching). Furthermore, buyer-side and seller-side A/B tests also achieve consistent results on the metrics that can be measured on either side. The counterfactual framework has been used for ongoing seller-side experiments on Marketplace and other products on Facebook.

ACKNOWLEDGMENTS

We would like to thanks McKenna Blenz, Wei Wu, Taroon Mandhana, Yanna Yan for their help and support on this work. Thank Corey Chen, Nade Sritanyaratana, Rongda Zhu, Srinath Aaleti, Yi-Hui Lin, Shalmoli Gupta, Qingyuan Kong, Neha Shah and Rachel Huang for fruitful discussions during the course of the project.

REFERENCES

[1] Deepak Agarwal, Bee-Chung Chen, Qi He, Zhenhao Hua, Guy Lebanon, Yiming Ma, Pannagadatta Shivaswamy, Hsiao-Ping Tseng, Jaewon Yang, and Liang Zhang. 2015. Personalizing LinkedIn Feed. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams (Eds.). ACM, 1651–1660. https://doi.org/10.1145/2783258.2788614

[2] Peter M. Aronow and Joel A. Middleton. 2013. A Class of Unbiased Estimators of the Average Treatment Effect in Randomized Experiments. *Journal of Causal Inference* 1, 1 (2013), 688–701.

[3] Peter C Austin and George Leckie. 2018. The effect of number of clusters and cluster size on statistical power and Type I error rates when testing random effects variance components in multilevel linear and logistic regression models. *Journal of Statistical Computation and Simulation* 88, 16 (Nov. 2018), 3151–3163.

[4] Lars Backstrom. 2016. Serving a Billion Personalized News Feeds. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*, Paul N. Bennett, Vanja Josifovski, Jennifer Neville, and Filip Radlinski (Eds.). ACM, 469. https://doi.org/10.1145/2835776.2835848

[5] Lars Backstrom and Jon M. Kleinberg. 2011. Network bucket testing. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar (Eds.). ACM, 615–624. https://doi.org/10.1145/1963405.1963492

[6] Sarah Bird, Krishnaram Kenthapadi, Emre Kiciman, and Margaret Mitchell. 2019. Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, J. Shane Culpepper, Alistair Moffat, Paul N. Bennett, and Kristina Lerman (Eds.). ACM, 834–835. https://doi.org/10.1145/3289600.3291383

[7] David Carmel, Elad Haramaty, Arnon Lazerson, and Liane Lewin-Eytan. 2020. Multi-Objective Ranking Optimization for Product Search Using Stochastic Label Aggregation. In *Proceedings of the International World Wide Web Conference (WWW), Taipei, Taiwan, April, 2020.*

[8] Na Dai, Milad Shokouhi, and Brian D. Davison. 2011. Learning to rank for freshness and relevance. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft (Eds.). ACM, 95–104. https://doi.org/10.1145/2009916.2009933

[9] Na Dai, Milad Shokouhi, and Brian D. Davison. 2011. Multi-objective optimization in learning to rank. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft (Eds.). ACM, 1241–1242. https://doi.org/10.1145/2009916.2010139

[10] Onkar Dalal, Srinivasan H. Sengamedu, and Subhajit Sanyal. 2012. Multi-objective ranking of comments on web. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, Alain Mille, Fabien L. Gandon, Jacques Misselis, Michael Rabinovich, and Steffen Staab (Eds.). ACM, 419–428. https://doi.org/10.1145/2187836.2187894

[11] Huan Gui, Ya Xu, Anmol Bhasin, and Jiawei Han. 2015. Network A/B Testing: From Sampling to Estimation. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi (Eds.). ACM, 399–409. https://doi.org/10.1145/2736277.2741081

[12] Viet Ha-Thuc, Srinath Aaleti, Rongda Zhu, Nade Sritanyaratana, and Corey Chen. 2019. Searching for Communities: a Facebook Way. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 1381–1382. https://doi.org/10.1145/3331184.3331426

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

[14] Ron Kohavi, Alex Deng, Roger Longbotham, and Ya Xu. 2014. Seven rules of thumb for web site experimenters. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, Sofus A. Macskassy, Claudia Perlich, Jure Leskovec, Wei Wang, and Rayid Ghani (Eds.). ACM, 1857–1866. https://doi.org/10.1145/2623330.2623341

[15] Jia Li, Dhruv Arya, Viet Ha-Thuc, and Shakti Sinha. 2016. How to Get Them a Dream Job?: Entity-Aware Features for Personalized Job Search Ranking. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (Eds.). ACM, 501–510. https://doi.org/10.1145/2939672.2939721

[16] Michinari Momma, Alireza Bagheri Garakani, and Yi Sun. 2019. Multi-objective Relevance Ranking. In *Proceedings of the SIGIR 2019 Workshop on eCommerce, co-located with the 42st International ACM SIGIR Conference on Research and Development in Information Retrieval, eCom@SIGIR 2019, Paris, France, July 25, 2019 (CEUR Workshop Proceedings)*, Jon Degenhardt, Surya Kallumadi, Utkarsh Porwal, and Andrew Trotman (Eds.), Vol. 2410. CEUR-WS.org. http://ceur-ws.org/Vol-2410/paper30.pdf

[17] Mario Rodríguez, Christian Posse, and Ethan Zhang. 2012. Multiple objective optimization in recommender systems. In *Sixth ACM Conference on Recommender Systems, RecSys '12, Dublin, Ireland, September 9-13, 2012*, Padraig Cunningham, Neil J. Hurley, Ido Guy, and Sarabjot Singh Anand (Eds.). ACM, 11–18. https://doi.org/10.1145/2365952.2365961

[18] Donald Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 5 (1974), 688–701.

[19] Donald Rubin. 1990. Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference* 25, 3 (July 1990), 279–292.

[20] Martin Saveski, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M. Airoldi. 2017. Detecting Network Effects: Randomizing Over Randomized Experiments. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*. ACM, 1027–1035. https://doi.org/10.1145/3097983.3098192

[21] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. 2002. Methods and metrics for cold-start recommendations. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, Kalervo Järvelin, Micheline Beaulieu, Ricardo A. Baeza-Yates, and Sung-Hyon Myaeng (Eds.). ACM, 253–260. https://doi.org/10.1145/564376.564421

[22] Krysta M. Svore, Maksims Volkovs, and Christopher J. C. Burges. 2011. Learning to rank with multiple objective functions. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar (Eds.). ACM, 367–376. https://doi.org/10.1145/1963405.1963459

[23] Liang Tang, Bo Long, Bee-Chung Chen, and Deepak Agarwal. 2016. An Empirical Study on Recommendation with Multiple Types of Feedback. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (Eds.). ACM, 283–292. https://doi.org/10.1145/2939672.2939690

[24] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon M. Kleinberg. 2013. Graph cluster randomization: network exposure to multiple universes. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, Inderjit S. Dhillon, Yehuda Koren, Rayid Ghani, Ted E. Senator, Paul Bradley, Rajesh Parekh, Jingrui He, Robert L. Grossman, and Ramasamy Uthurusamy (Eds.). ACM, 329–337. https://doi.org/10.1145/2487575.2487695