A Cross-Platform A/B Testing Framework for Offsite Advertising

Shichuan Ma
Indeed.com
Sunnyvale, CA, USA
shichuanm@indeed.com

Fengdan Wan
Indeed.com
Sunnyvale, CA, USA
fwan@indeed.com

Ziying Liu
Indeed.com
Sunnyvale, CA, USA
ziyingl@indeed.com

Yu Sun
Indeed.com
Sunnyvale, CA, USA
sunyu@indeed.com

Haiyan Luo
Indeed.com
Sunnyvale, CA, USA
hluo@indeed.com

ABSTRACT

A/B testing is a tool that has been widely used in the industry for product performance measurements and optimizations. In our marketplace growth product, we leverage various 3rd party vendor platforms to display our job ads in the open Internet. In this paper, we described an innovative approach to conduct A/B testing across various vendor platforms. This A/B testing framework allows internal stakeholders to configure A/B tests through a well-developed user interface (UI) without code changes. It automatically populates the A/B split to various vendor platforms as desired. Since launch, this A/B testing framework has been widely used in our product and provides a unique way for us to fine-tune ads performance holistically across those vendor platforms.

KEYWORDS

A/B testing, online evaluation, view-through

ACM Reference Format:
Shichuan Ma, Fengdan Wan, Ziying Liu, Yu Sun, and Haiyan Luo. 2022. A Cross-Platform A/B Testing Framework for Offsite Advertising. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Indeed.com is the world No.1 job search website. With millions of jobs posted online everyday, indeed has attracted tons of job seekers daily. In a hot job market, where sometimes the number of job openings is significantly larger than active job seekers, we want to expand the visibility of job postings and company awareness to a broader population. We also want to re-engage with job seekers while they are not on indeed.com. Due to these two reasons, we built a marketplace platform on which Indeed job postings can be displayed on 3rd party websites. This technique is called offsite advertising.

One of the biggest challenges for the marketplace platform is to optimize the ads performance holistically with multiple integrated vendors. We have been working with multiple publishers, including walled gardens such as Google and Facebook, and demand-side platforms (DSP) such as Beeswax. They provided various ads features that may influence the performance of the ad campaigns. We need to tune each feature in order to maximize the performance.

A/B testing, also known as split testing, is widely used to optimize online implementations [9][12][11][4][8]. By showing two versions of a design or an implementation to two different groups of users, A/B testing is helpful to find which version yields better performance. With the fast development of online advertising, A/B testing is also widely used to improve the performance of online advertisements. For example, an effective method was developed in [2] to estimate the causal effect of the marketing campaigns. In the data driven world, vast majority of Internet companies are leveraging A/B testing to measure multiple variants performance. Indeed has its own open sourced A/B testing framework [7].

Most of the ad platforms, such as Facebook [5] and Beeswax [3], also support A/B testing for online advertising. With simple configurations, advertisers can set up A/B tests of selected features and specify how the users are split into the control and test groups. However, one job seeker could be put into the control group on one platform, but into the test group on the other platform. This issue may cause ambiguity in the performance analysis of the A/B tests across different platforms. If, for example, one user is grouped into the control group by one ad platform for one A/B test, this user doesn't see the test variant on this platform. But the same user could be grouped into the test group of the same A/B test by the other platform, leading to exposure of the test variant to this user. Therefore, it is not possible to determine whether and how the performance changes related to this specific user is caused by the A/B test. None of the aforementioned research works discuss or resolve such a cross-platform user splitting issue.

In this paper, we propose an A/B testing framework that can work across multiple ad platforms. By strictly defining the control group, this framework sets up A/B tests on different ad platforms but keeps a selected portion of users in the control group regardless of the platforms. This design prevents users in the control group from seeing the test variants in any supported ad platforms.

It's worth noting that the framework does not restrict a uniformed user split across all ad platforms. Instead, the user split can also be platform specific for experiments that can be conducted orthogonally on ad platforms.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Section 2 describes the main methodology of the A/B testing framework followed by some implementation remarks in Section 3. Section 4 illustrates 4 experiments that were conducted via this A/B testing framework. We concludes the paper in Section 5.

2 METHODOLOGY

A basic problem of building a generic A/B testing framework that can work across multiple ad platforms is how to split a selected subset of all reachable users into the control group regardless of the ad platforms. Since ad platforms don't have knowledge of each other, the solution must be found at Indeed's side. We propose to use the audience re-targeting, which is supported by all ad platforms, to limit the users in the control group of A/B tests.

Audience targeting [6] is a technique widely used by online advertisers to show ads only to the people who are most likely to be interested in the ads. One common way to implement the audience targeting is to categorize the users into different segments, and use the segments to match the ads. For example, at Indeed, the job seekers, who are nurses as mentioned in their resume or are searching for nursing related jobs, could be put into the nursing segment, and would likely be targeted for ads that target the nursing segment.

In order to let the ad platforms use Indeed's segments, we need synchronize the job seekers at Indeed with the ad platforms. This can be done by cookie synchronization, a technique to map user IDs from one system to another [10].

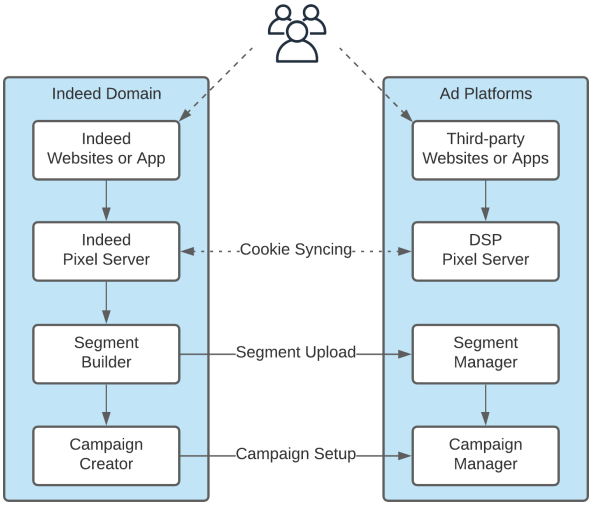

Figure 1: Block diagram of the offsite platform

Fig. 1 is a block diagram of the marketplace growth platform. A job seeker who visited Indeed.com or used Indeed mobile app will be assigned to a unique user id. At Indeed, we use cookie tracking key (CTK) as the unique user id. By examining the user's profile and recent activities, such as search and clicks, we can categorize the user to a list of segments.

Similarly, the job seeker will be assigned a unique id by an ad platform when he or she visits the ad platform. Different ad platforms may use different formats for the id. For example, Facebook uses email while Beeswax uses a unique string. Let's use *userId* to represent the unique user id in an ad platform. By using cookie synchronization, we can obtain the mapping from the *CTK* to the *userId*.

In this way, we obtain the mapping from *CTK*s to *userId*s for all Indeed job seekers. We then upload audience segments to the data management platform (DMP) of the ad platforms. This procedure again varies from one ad platform to another, but the concept is similar. Two steps are usually involved. First, create a segment in the ad platform. A unique key is usually required. We use *Indeed-* < *segmentId* > as the key, where "Indeed-" is a prefix and *segmentId* is the internal id of the segment. Second, upload the list of the *userId*s and link them with the segment key. After finishing this step, we have shared the user ids and have created segments in the ad platforms. We are now ready to set up offsite campaigns.

Most ad platforms provide both UI and API to create campaigns. The procedures are similar to each other. One campaign needs to be created first. A campaign objective, a budget, and a campaign running time can be specified at the campaign level. Multiple ads can then be created under the campaign. Many features can be specified at the ad level, such as budget, pacing strategy, bidding strategy, frequency cap, and targeting. Next, one or more creatives can be created and associated with an ad. One creative defines how the ad looks like and how to respond to user activities, such as views and clicks. We don't cover the details of the campaign setup since it isn't the focus of this paper.

As we are using audience targeting to specify the users in both control and test groups, and the audience targeting can be only applied at the ad level, we can only perform the A/B test at the ad level, but not at the campaign level. This limitation is acceptable because most of the tested variants are not related to the features at the campaign level.

At this point, we have successfully set up test group ads and control group ads respectively. The impressions and clicks events are collected in respect to the control and test groups. We performed power analysis towards those data sets to evaluate the performance metrics from those groups.

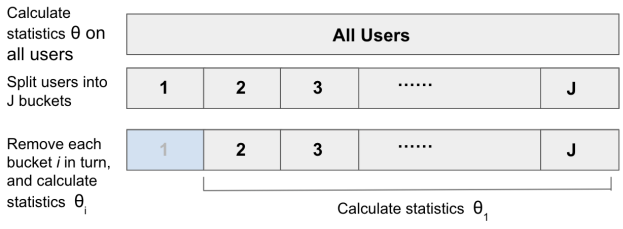

Figure 2: Jackknife Algorithm for Confidence Interval Estimation

To estimate confidence interval in our experiments, we leveraged the jackknife algorithm. Fig.2 demonstrates the Jackknife algorithm for confidence interval estimation. For a given statistics θ whose

confidence interval we are trying to estimate on population P, the algorithm starts with calculating statistics θ over the whole population. We then split the population into J non-overlapping groups, followed by removing each bucket i in turn, and calculating our statistics θ_i for the remaining population. Finally, the standard error e for statistics θ is estimated using the following formula:

$$e = \sqrt{\frac{J-1}{J}\sum_{j=1}^{J}(\hat{\theta} - \theta_i)^2}$$

where

$$\hat{\theta} = \frac{\sum_{j=1}^{J}(\theta_j)}{J}$$

We eventually report the confidence interval for statistics θ as

$$[\theta - t_{J-1,0.975}e, \theta + t_{J-1,0.975}e]$$

Here t is the 97.5th percentile of the T Distribution with $J-1$ degrees of freedom.

3 IMPLEMENTATION

We have implemented the proposed A/B testing framework in our marketplace growth product. Addition to the requirement that the same job seekers are put into the control group across ad platforms, we add more requirements in the implementation as listed below.

- Support flexible ramping up or down of the test group to different percentages
- Support configurable inputs of Indeed onsite ads
- Support pre-defined A/B tests in order to avoid code changes

To simplify our implementation and to support aforementioned features, we utilize Proctor, Indeed's open source A/B testing framework [7], as the base framework of the implementation. As an example, Fig.3 shows the Proctor configuration of one creative A/B test.

As one of the major functionalities, Proctor helps to allocate users to different buckets. One bucket represents one user group. This user allocation can be done dynamically, which means the percentage of users allocated to each bucket can be adjusted in the UI of the Proctor without code change. As shown in Fig.3, we configured a 50-50 test for the creative A/B test. We configure the A/B test by adding variables in the Constants section. In this example, we defined the partner as Beeswax, the onsite sub product as ITAB (Indeed Targeted Ads - Brand), and specify a list of campaign IDs. We support more configurations, such as various input methods, start/end time, and automatic stop conditions. Due to the page limitation, we don't cover all configurations in this example.

We define the A/B test content by using the Json payload of each bucket. In the example, we leave the payload of the ctrl bucket empty, making the control line as the baseline (no change). The payload of the test bucket is set to remove creative template V1 and add creative template V2. Therefore, the purpose of this A/B test is to compare the performance of the creative template V2 against V1. We have pre-defined a list of tests to cover all popular A/B tests in our project and implemented the code changes, leading to a config-n-test environment.

We also added offsite_ab_tests as one of the meta tags. It is used by the segment generator/uploader to filter offsite A/B tests from all company-wide Proctor tests. The procedure is listed in Alg.1.

Algorithm 1 Generate and upload A/B test segments

```
Get a list of Proctor tests with meta tag offsite_ab_tests
for each Proctor test with id testId do
    Create a segment for each bucket (bucketId)
    Name the segment Indeed-testId-bucketId
    for each user do
        Call the Proctor API to get the allocation for this user
        Link the associated userId to the corresponding segment
    end for
end for
Create all segments in ad platforms
Upload the mapping from segment to a list of userIds
```

After uploading the segments, we can set up A/B tests. The procedure is listed in Alg.2.

Algorithm 2 Set up A/B tests

```
for each A/B test with id testId do
    for each ad in the list of campaign_ids do
        Assume the ad is targeting an audience segment Indeed-segmentId
        for each bucket with id bucketId do
            Duplicate the ad in the ad platform
            Change the audience targeting of the baseline to (Indeed-segmentId & Indeed-testId-bucketId), where (id1 & id2) denotes job seekers in both segment id1 and id2
            Apply the changes defined in the payload
        end for
    end for
end for
```

4 EXPERIMENTS

Leveraging the A/B testing framework, we have conducted several A/B tests. These experiments have helped us fine-tuned campaign settings, analyzed offsite ads performance and boosted the offsite ads performance significantly. We share three experiments in this section.

4.1 View through analysis

Compared to Indeed onsite precious inventory, job ads tend to not perform well in the open Internet websites directly. For example, one job ad with onsite CTR of 2.4% can easily drop its CTR to around 0.5% offsite. The reason is simply that indeed.com is known as a job hunting website and all job seekers visiting indeed.com are in the mode of finding some job information. However, when they are browsing other sites, they might not in the job hunting mode and are hence less likely to respond to job related ads. The purpose of delivering job ads offsite is not only improving its direct

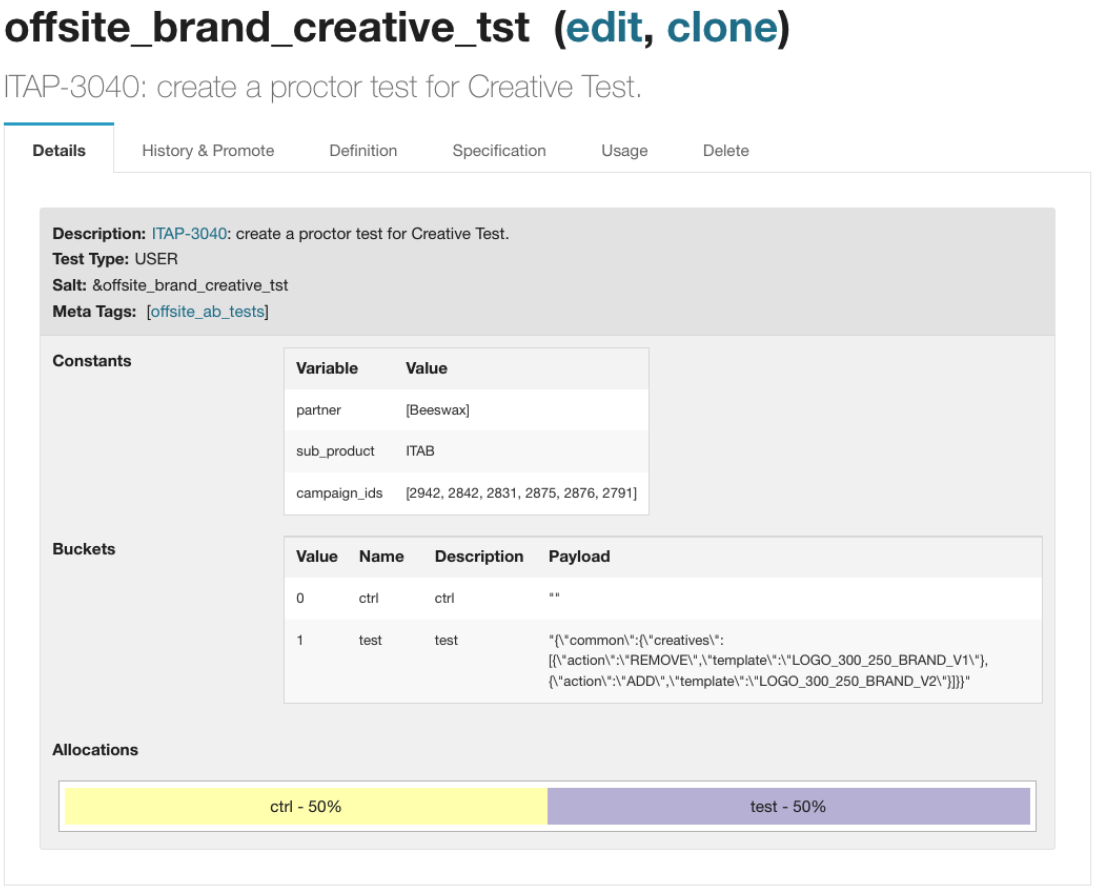

Figure 3: An example of Proctor configuration

performance metrics, such as cost per apply and cost per click, but more strategically, improving overall job seeker's engagement and improving indirect performance. Our hypothesis is that ads impressions are valuable. With more impressions shown to job seekers while they are browsing the Internet, their overall engagement with Indeed, the company and jobs will be enhanced, hence help advertisers and Indeed yield better ROI.

To test the hypothesis and evaluate offsite's ads value propositions, we proposed to measure offsite ads' impression values through so-called view through attribution analysis. We designed an experiment where 10% Indeed audiences (defined by CTK) are in the held off group and 90% of audiences in the active bucket. The held-off group will see a generic offsite campaign, called Public Service Announcement [1] lines, and the active group will see the actual offsite ads, such as job ads.

The platform collected offsite view-able impression events and joined them with users' onsite activities. The analysis generates various meaningful metrics to compare users' onsite activities between held-off group and active group. Table 1 shows job seeker onsite activities after seeing the offsite branded ads within 1-, 7-, 15- and 30-days windows. The values in the table illustrate the lift of onsite activities from active group compared to the held-off group.

Table 1: branded ads per job seeker (js) engagement incremental lift

	1 day	< 7 days	< 15 days	< 30days
onsite visits per js	4.67%	4.44%	3.33%	1.47%
applications per js	12.12%	9.57%	7.16%	4.71%
apply starts per js	11.49%	9.45%	7.23%	5.54%
job clicks per js	6.42%	6.55%	4.76%	3.79%
job searches per js	2.64%	5.58%	3.62%	1.51%

All numbers here are statistically significant. We can see that job seekers engaged more with Indeed after viewing the offsite ads. Furthermore, the lift is more obvious within the first week after ads are shown.

Table 2 shows the metrics of advertising company. It proved the offsite ads values for the advertising companies. As similar to the previous table, the incremental lift has been computed for the 1-day, 7-days, 15-days and 30-days windows. All the above results are statistically significant. The above results showed that in general offsite impressions improved overall job seeker's engagement on Indeed.com, positive outcomes of the advertising companies, as well as helped the revenue growth. Noticeably taken away over the

Table 2: branded ads - Advertising company per job seeker (js) metrics

	1 day	< 7 days	< 15 days	< 30days
Views of company page for advertising company per js	1406.77%	1993.29%	239.01%	172.53%
Views of job for advertising company per js	42.01%	52.58%	66.67%	81.10%
Apply starts for jobs for advertising company per js	12.89%	11.25%	18.78%	30.41%
Company searches for any company per js	12.18%	14.06%	8.11%	6.87%
Views of company page for any company per js	9.54%	3.04%	1.73%	-0.21%

Table 3: job ads per job seeker (js) engagement incremental lift

	1 day	< 7 days	< 30days
onsite visits per js	-0.35%	-0.28%	2.36%
applications per js	2.83%	1.53%	7.27%
apply starts per js	2.51%	0.51%	7.21%
job clicks per js	3.85%	0.42%	6.11%
job searches per js	0.84%	-10.14%	3.95%

course of time, the effectiveness of offsite impressions diminishes as expected but there are also outliers in this trend. One possible explanation for this is offsite ads have been displayed to the audience multiple times during experiments, so there could be an overlapped period in the measurement. Despite this caveat, the results discussed above truly called out offsite ads' values, because hard to control is considered as one of offsite environment characteristics.

We do not get statistically significant data per advertising companies in these categories. As noted, job ads performance is not significantly impactful as compared to brand ads. There are a lot of factors leading to this, one of them is specific job ads, user targeting is critical. For branding campaign, user targeting can be loose. Due to increasing high volumes of job ads from job categories targeting for specific audience, data is sparse and we did not gather sufficient data in several dimensions. Form key matrices listed in Tab.3, the results demonstrates the impressions values of offsite ads.

There are some other key metrics that the company is interested in are ongoing analysis now. By a quick glance, the offsite ads increased number of account creations, resume uploads significantly as well. All the metrics above demonstrated the offsite ads value propositions.

4.2 Bidding strategy test

We leveraged the A/B testing framework to test out various bidding strategies. This helped us fine tune the campaign optimal bidding in term of CTR, the main metric we used here. Our default bidding strategy was flat CPM bidding. We experimented with $3 CPC bidding strategy. The testing group showed positive on all fronts, in terms of reach, CTR, and vCTR. CTR is considered as one of key metrics we measured. Based on the test results, we rolled out the $3 CPC bidding strategy as our baseline (default setting).

4.3 Lookalike segmentation testing

We also tested out our lookalike algorithm as audience expansion to increase reach via the A/B testing framework. The results show increased impressions by 130+%, increased reach to 2.3X. With this result, we can safely expand the segments scoring to include those audiences into the offsite audience segments.

5 CONCLUSION

This paper illustrated the A/B testing framework we built in our marketplace growth platform. The framework is generic to work across 3rd party platforms. We have encountered engineering challenges while implementation, for example, how to effectively troubleshoot and debug since the configuration shall be populated to all 3rd party platforms and setup A/B groups there correctly. There are also scientific challenging, e.g., how to split users groups to conduct statistically significant tests, etc. Nevertheless, this framework has helped prove the offsite product values propositions and guide us on making decisions. With this framework in place, we can demonstrate and improve our marketplace growth product efficiently.

ACKNOWLEDGMENTS

We want to thank all team members of the Ads System team for their contribution to this product.

REFERENCES

[1] 2021. public service announcement. https://en.wikipedia.org/wiki/Public_service_announcement
[2] Joel Barajas, Jaimie Kwon, R. Akella, Aaron Flores, Marius Holtan, and Victor Andrei. 2012. Marketing campaign evaluation in targeted display advertising. (08 2012). https://doi.org/10.1145/2351356.2351361
[3] Beeswax. December 16, 2021. Beeswax Experiments. https://docs.beeswax.com/docs/experiments
[4] David Chan, Rong Ge, Ori Gershony, Tim Hesterberg, and Diane Lambert. 2010. Evaluating online ad campaigns in a pipeline: Causal models at scale. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 7–16. https://doi.org/10.1145/1835804.1835809
[5] Facebook. 2021. About A/B Testing. https://www.facebook.com/business/help/1738164643098669?id=445653312788501
[6] Google. 2021. About audience targeting. https://support.google.com/google-ads/answer/2497941?hl=en
[7] Jack Humphrey. 2014. Proctor: Indeed's A/B Testing Framework. Retrieved December 16, 2021 from https://engineering.indeedblog.com/blog/2014/06/proctor-a-b-testing-framework/
[8] Ron Kohavi, Thomas Crook, Roger Longbotham, Brian Frasca, Randy Henne, Juan Lavista Ferres, and Tamir Melamed. 2009. Online Experimentation at Microsoft. (09 2009).
[9] Ron Kohavi and Roger Longbotham. 2017. *Online Controlled Experiments and A/B Testing*. Springer US, Boston, MA, 922–929. https://doi.org/10.1007/978-1-4899-7687-1_891
[10] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P. Markatos. 2018. Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask. *CoRR* abs/1805.10505 (2018). arXiv:1805.10505 http://arxiv.org/abs/1805.10505
[11] D. Siroker and P. Koomen. 2013. *A / B Testing: The Most Powerful Way to Turn Clicks Into Customers*. Wiley. https://books.google.com/books?id=VfVvAAAAQBAJ
[12] Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. 2015. From Infrastructure to Culture: A/B Testing Challenges in Large Scale Social Networks. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015).