A Unified Framework for Evaluating Online User Treatment Effectiveness, with Advertising Applications

Pengyuan Wang[†] Marsha Meytlis[†] Fei Yu [§] Jian Yang[†]
[†]Yahoo Labs, Sunnyvale, CA, USA
[§]Purdue University, West Lafayette, IN,USA
[†]{pengyuan, mmeytlis, jianyang}@yahoo-inc.com [§]feiy@stat.cmu.edu

ABSTRACT

The measurement of ad effectiveness is one of the central problems of online advertising. Typically the performance is measured by investigating the proportion of people who converted or performed other success actions after they saw the ads. These metrics commonly overestimate campaign effectiveness since they do not account for users who would have performed actions even if the campaign did not happen. Conventional metrics also fail to answer the following questions that are important to advertisers: 1) Which users convert because they see the ad and which users would have converted even if they do not see the ad? 2) What is the cumulative effect of multiple advertising strategies on performance? 3) How does a campaign affect the size of the potential audience pool?

In this paper we propose a general methodology for assessing campaign performance that addresses all of these questions. Our method does not require randomized experiments or additional ads to be shown. We develop a unified causal modeling framework that establishes a causal relationship between seeing an ad and performing an action, which is based on propensity methodology. We derive a novel robust rank test for model validation. We also provide innovative interpretations of the estimation results by the causal inference, addressing 'smart cheating' of online ads (i.e. targeting the users who are likely to convert even without any ad exposure, which does not add value to the advertisers). The three components (model, validation, and interpretation) complete a unified solution to ad effectiveness measurement. The framework is applied to three online campaigns involving millions of unique users. Results from real online campaigns show that this methodology is robust to online data sparseness, high dimensionality and biases from user features.

This paper focuses on measuring the effectiveness of online ads, but the framework is readily applicable to measure the effectiveness of other kinds of treatments on various user metrics, for example the impact of different strategies on user engagement metrics.

Categories and Subject Descriptors

G.3 [**PROBABILITY AND STATISTICS**]: Statistical Computing; J.1 [**ADMINISTRATIVE DATA PROCESSING**]: Business, Marketing

Keywords

Advertising, Causal Inference, Propensity Score, Robust Rank Test, Smart Cheating

1. INTRODUCTION

The measurement of ad effectiveness is one of the central problems of online advertising. It is important to be able to determine whether an advertising campaign leads to better performance or not. Typically the success rate[1] of a campaign is measured as the percentage of users who complete a certain desired action; however, this metric does not provide a complete assessment of performance since 1) it does not account for users who would have performed actions even if the campaign did not happen, and 2) the measure of ad effectiveness has multiple dimensions.

We develop a new method for measuring campaign performance that evaluates three dimensions: 1) the direct effect of a single advertising strategy on user performance 2) the effect of multiple advertising strategies on user performance, and 3) the effect of ad campaign on audience pool expansion. We propose a unified pipeline to measure all three dimensions of performance.

1. The first dimension is a metric that measures uplift of a single ad placement [31, 19, 10]. Uplift is a metric that measures change in online brand interest that results from additional users who are recruited by a campaign. Users who perform regardless of whether they see an ad need to be discounted, which requires unbiased estimations of the portion of users who will convert without ad exposure.
2. The second dimension of our analysis is a metric that measures the cumulative effect of multiple campaigns on user performance [6, 24, 10]. Frequently an advertiser may run several campaigns simultaneously, such as a website takeover and a mobile campaign, or a video campaign and a direct response campaign. The advertiser needs to not only know the uplift of each individual campaign, but also how each of these campaigns enhance one another. Our analysis assesses the cumulative effect of these campaigns, and we call this cumulative effect: synergy.
3. The third dimension of our assessment is to determine how the potential customer pool changes as a result of the campaign. Typically customers need to show brand awareness before they are ready to make a commitment to purchase a product. This process of learning about a product and then deciding to buy the product is referred to as traveling down the purchase funnel. Our analysis gives insight on how many new users have entered the purchase funnel because of learning about the product in a branding ad campaign.

All the three dimensions described require a fair comparison of the responses to difference advertising treatments. For example, the uplift assessment requires a comparison of the success rates of the people who saw (exposed group) and not saw the ads (control group, i.e. non-exposed group); and synergy assessment requires

[1] A success or conversion performance is an action favored by the campaign, such as click, search or site visitation. Success rate is the percentage of unique users who take a success action. In this paper we use success and conversion interchangeably.

a comparison of the success rates of the people who saw ads from both placements (exposed group) and a single one (control group). One method to obtain the non-biased assessments of the success rates, besides our proposed model, is a randomized experiment, i.e., an A/B test. The success rates of the two groups are unbiased in an ideal AB test, because the exposed and control users are randomly picked from the same audience and have the same characteristics. However a randomized test may not always be available, and in an observational advertising campaign the direct comparison between exposed and control may be biased if control users have different features than the exposed users. Here's an example that illustrates this bias. Imagine a cosmetic product campaign where all of exposed users are females and all of the control users are males. If the females generally have a larger conversion rate than males, the uplift (effectiveness) of the campaign could be overestimated because of the confounding effect of the user features, in this case, gender. In such cases, the high success rate of the exposed group is not caused by ads, and hence cannot serve as a fair measurement of ad effectiveness. In order to establish a causal relationship between ad treatments and conversions, such biases from user features need to be eliminated. The intuition behind this argument is illustrated in Figure 1, where the ad effect on conversion is confounded by the features (which is gender in this example). One needs to eliminate the impact of features as in Figure 2 to isolate the real causal impact of ads on conversions.

An immediate attempt to eliminate the impact of user features is to estimate individual performance (e.g. conversion or no conversion) using a regression model. Such a model would estimate the relationship between individual performance (dependent variable) and independent variables, where the independent variables would consist of user features and ad exposure indicators. However it is well known that correlation does not imply causation. Such conventional performance model fails to make the distinction that temporal correlation between ad exposure and performance does not imply causation. Hence the causal effect of the ad exposure is difficult to estimate by adjusting the outcome with the user features directly, which is further shown as in Section 3.4.

To address the problem of biases in ad effectiveness assessment in the above-mentioned conventional methods, we develop a novel statistical approach that can be used on online data to address the three dimensions of performance discussed above. Our methodology is based on a causal model that balances the user features of the exposed and control groups, and hence establishes a cause and effect relationship between seeing an ad and performing actions. The causal model enables us to measure the three aspects of ad effectiveness (uplift, synergy, and audience pool expansion) in a unified framework.

Our causal model is based on a statistical matching approach utilizing inverse propensity weighting (IPW)[27] and doubly robust (DR) estimation [25, 30]. Defining the two ad treatments as 'control' and 'exposed', the propensity score is defined as the estimated probability for a subject to be exposed, given a set of observed features or pre-treatment covariates. Typical methods to estimate propensity scores include linear logistic regression [28, 29], semiparametric regression [17], and non-parametereic regression [11, 22]. More robust results are reached via DR estimator [25, 30], which is proven to have a smaller asymptotic variance. This method rapidly becomes popular in various fields, including economics [12], health care [3], social science [16], politics [14], online behavior study [9] and advertising [5, 18, 32, 8, 2]. In the online advertising area, Chan et al. [5] considered the industrial advertising data, with moderate size. [18] showed that observational data may lead to incorrect estimates, [32] explored the benefits of estimating several other parameters of interest and another method targeted maximum likelihood estimation (TMLE)),[8] used causal inference for a multi-attribution problem, and[2] used it in an experimental circumstance.

None of the methods proposed before were used to construct a full solution to measure ad effectiveness. In addition, the previous propensity-based methods have not been tested on real live campaigns involving millions of users or addressed the sparsity, huge volume and large number of user features, which typically exist with online advertising data.

In our previous publication [35], we proposes a propensity-based framework that addresses the above problems, but it is lack of two components: a model validation approach that is robust to outliers and skewness of the data, and an approach to interpret the result from business point of view. Also it was not used to propose a full solution to measure ad effectiveness, including uplift, synergy and audience pool expansion aspect. In this paper, we devise a unified pipeline to measure the three aspects of ad effectiveness, and the modeling framework is divided into the following modules.

1. Model: The propensity-based causal inference framework to address the sparsity and huge volume in industrial datasets (Section 2). The framework is first described in [35], and we briefly introduce it here as a background.
2. Validation: An innovative robust rank test for model validation (Section 3).

   In order to validate our model, we need to check that the propensity-based weighting method has balanced the control and exposed groups. The standard approach to check the balancing effect of the weighting is the standardized mean difference, i.e. the two sample weighted t-test [26, 15, 21]. This approach has been applied in various areas, e.g. politics [14], public health [23], finance and management science [34] and media [20], etc. However this test is vulnerable to the skewness and outliers in covariate distributions, which is often the case in advertising data. To address the non-robustness in the model verification of the traditional method, we devise a novel robust rank test for covariate balancing verification. The advantage of the proposed method is verified with simulation. To the best of our knowledge, we are the first to address the skewness of advertising data with a robust weighted rank test.
3. Application: The above approaches are applied to three online campaign involving millions of unique users, to assess three aspects of ad effectiveness: uplift, synergy, and audience pool expansion (Section 4).
4. Interpretation: Novel interpretations of the results, addressing 'smart cheating' of online ads (Section 5).

   A major concern for online advertising is that, some of the users might convert even without any ad exposure. Targeting on this part of users might result in high conversion rates but actually does not add to the value of the advertisers. We devise a strategy to interpret the calculation result, which reveals the 'smart cheating' or the 'honest reaching' in ad placements.

Our framework is not limited to online advertising, but is also applicable to other circumstances (e.g., social science) where causal connection needs to be built with observational data.

2. BACKGROUND

In this section we briefly review the causal inference framework as in [35], which addresses the sparsity, huge volume, and large amount of features in real live campaigns.

The causal effect is measured by comparing a specific treatment

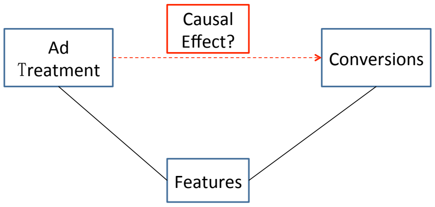


Figure 1: Confounding Effect of Features

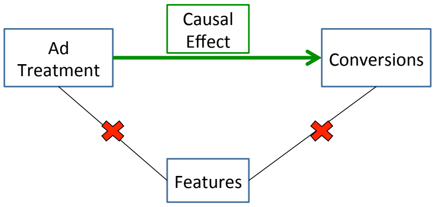


Figure 2: Confounding Effect of Features Eliminated

(exposure) toward another (control). For example in the case where we measure the effect of a website takeover in conjunction with a targeted display campaign, 'exposure' treatment is impressions from both placements while 'control' treatment is impressions only from the targeted display placement. Each subject has two potential outcomes Y_c and Y_e under control and exposure treatments, respectively. However we only observe a single outcome for a given subject.

Suppose that the control (e.g. no ad impression) and exposure (e.g. ad impressions) treatments are applied to two groups of subjects with no overlaps, indicted by $z_i = 0$ (control group) or 1 (exposed group) for subject $i = 1, 2, ..., N$. The success metric (such as conversion), is indicated by $y_i = 0$ or 1. A naive way to estimate $E(Y_e)$ and $E(Y_c)$ is to calculate the average success rates of the two groups, respectively, as in Equations 1 and 2.

$$r_{\mathrm{naive},exposed} = \frac{1}{\sum_i z_i} \sum_i z_i y_i; \tag{1}$$

$$r_{\mathrm{naive},control} = \frac{1}{\sum_i (1-z_i)} \sum_i (1-z_i) y_i. \tag{2}$$

The difference or ratio of $r_{\mathrm{naive},exposed}$ and $r_{\mathrm{naive},control}$ can be used to evaluate the effectiveness of exposure, as in Equations 3 and 4. In the rest of the paper we will use amplifier to indicate the ratio of the conversion rates of the exposed and control groups.

$$D_{\mathrm{naive}} = r_{\mathrm{naive},exposed} - r_{\mathrm{naive},control}; \tag{3}$$

$$R_{\mathrm{naive}} = r_{\mathrm{naive},exposed} / r_{\mathrm{naive},control}. \tag{4}$$

The naive estimators are unbiased if the control and exposed groups of users are randomly sampled from the population. However in observational studies, the ad treatments might be highly related to user features, such as network activity, website visitation, demographics, etc. In such cases, the estimated $r_{\mathrm{naive},exposed}$ and $r_{\mathrm{naive},control}$ cannot represent the whole population, and hence it does not ensure the comparison of the two groups is on equal footing.

A straightforward approach to eliminate the impact of user features on the outcome, is to adjust the outcomes with a set of user-level covariates X_i as well as treatment indicators, i.e. fit a model $y_i \sim f(X_i, z_i)$. However the estimated effect of z_i may not necessarily imply causal effect, since the model may not correctly address the relationship between z_i and X_i. This is further discussed in Section 3.4.

A sound approach to address the different features of the control and exposed groups is to use the IPW, which considers the treatment z_i as a random variable depending on a set of pre-treatment covariates X_i for each subject i.

We define propensity score as the probability $\hat{p}_i = P(z_i = 1|X_i), \forall i$, whose estimator $\hat{p}_i$ is usually obtained by fitting a model $\hat{P}(X)$ to estimate probability to be exposed with respect to the covariate X. Specifically we model $\hat{p}_i \sim \hat{P}(X_i)$ where $z_i = 1$ with probability $\hat{p}_i$. The basic idea is to use the estimated $\hat{p}_i$ to match the control and exposed groups, rather than to match the multi-dimensional X.

The IPW method proposes that each control subject is weighted by $1/(1-\hat{p}_i)$ and the exposed subjects is weighted by $1/\hat{p}_i$. [2] Hence the weighted success rates of the control and exposed groups are defined as in Equations 5 and 6.

$$r_{ipw,exposed} = \frac{1}{\sum_i 1/\hat{p}_i} \sum_i z_i y_i / \hat{p}_i; \tag{5}$$

$$r_{ipw,control} = \frac{1}{\sum_i 1/(1-\hat{p}_i)} \sum_i (1-z_i) y_i / (1-\hat{p}_i). \tag{6}$$

The IPW estimation of ad effectiveness (difference and amplifier) are then defined as the difference and ratio of $r_{ipw,exposed}$ and $r_{ipw,control}$ respectively.

With proper assumptions [27][3], the IPW is proved to be unbiased.

The above estimator measures the average exposure effect over the whole population. In practice, we may also be interested in the average exposure effect on the subpopulation of subjects who actually got exposed, which is called the treatment on treated effect (TTE). For this estimation, the control subjects are weighted by $\hat{p}_i/(1-\hat{p}_i)$ and the exposed subjects are not weighted, as in Equations 7 and 8. Hence the ad effectiveness is measure by the difference or ratio (amplifier) of $r_{ipw,tte,exposed}$ and $r_{ipw,tte,control}$.

$$r_{ipw,tte,exposed} = \frac{1}{\sum_i z_i} \sum_i z_i y_i; \tag{7}$$

$$r_{ipw,tte,control} = \frac{1}{\sum_i (1-z_i)\hat{p}_i/(1-\hat{p}_i)} \sum_i (1-z_i) y_i \hat{p}_i / (1-\hat{p}_i). \tag{8}$$

In this paper we choose to use GBDT to model the propensity score ($\hat{P}(X)$) and success probability under control ($\hat{M}_0(X)$) and exposure ($\hat{M}_1(X)$) treatments with covariate X. We compared GBDT with several popular methods, including PCA [13] for feature selection, and logistic regression, LASSO [33] and random forest [4], and verifies that GBDT outperforms these methods and provide reasonably good estimations. Note the choice of propensity model is not the focus of this paper. To improve the computation efficiency, it is possible to choose more scalable models and algorithms, for example logistics regression propensity models and the scalable algorithms in [1].

Also to deal with the large volumes of users, we divide the whole dataset into subsamples, and conduct the analysis within each data chunk. The computation on each of the subsamples yields an estimation of ad effectiveness, and the point estimation and variation of the population-level ad effectiveness are summarized from the collected subsample estimations.

[2] The basic intuition is that, a control subject belongs to its group with probability $1-\hat{p}_i$, and hence it is weighted by the inverse of this probability to infer the situation of the population. Similarly for the exposed subjects. See [25] for proofs.

[3] Assumption 1: Stable unit treatment value assumption. "The (potential outcome) observation on one unit should be unaffected by the particular assignment of treatments to the other units" [7]. Assumption 2: Strong ignorability of treatment assignment (also called "Unconfoundedness")[27]. Given the covariates X, the distribution of treatment assignments is independent of the potential outcomes.

3. MODEL VALIDATION WITH RANK TEST

As mentioned in Section 1, the propensity-based weighting method in Section 2 is aiming to balance the control and exposed groups. The conventional standardized mean difference is not robust to skewness in covariate distributions. In this section we proposed a novel weighted rank test for this task, which completes the framework with robust covariate balancing effect checking.

Suppose each of the users are assigned weight w_i according to IPW or IPW the TTE estimator. The conventional method utilizes the test statistic $\frac{\mu_{exposed}-\mu_{control}}{\sqrt{\frac{\sigma^2_{exposed}}{\sum_i z_i}+\frac{\sigma^2_{control}}{\sum_i(1-z_i)}}} \sim N(0,1)$, for each feature h with observed covariate $x_{i,h}$,

where $\mu_{exposed} = \frac{1}{\sum_i z_i w_i}\sum_i z_i x_{i,h} w_i$, $\sigma^2_{exposed} = \frac{1}{\sum_i z_i w_i}\sum_i z_i x^2_{i,h} w_i - \mu^2_{exposed}$ and $\mu_{control}$ and $\sigma^2_{control}$ are computed similarly.

However, the standardized mean difference test is vulnerable to heavy-tail distributed features and outliers. In advertising dataset, the user activity and features are typically heavy-tail distributed, which can be seen in Figure 7 (a) (b). In this paper, we proposed a weighted Mann-Whitney-Wilcoxon rank test to deal with the heavy-tailness of the observed dataset.

The Mann-Whitney-Wilcoxon rank test [21, 36] is a nonparametric test for checking whether a sample is stochastically larger than another sample. It is known that the Mann-Whitney-Wilcoxon rank test does not assume any specific form for the distribution of the population and hence is more robust when the underlying distribution is not normal. The original version of the MWW test is developed in multiple ways, for example in [37]. In the causal inference framework, each observation is weighted according to its propensity score. We derive a weighted version of the Mann-Whitney-Wilcoxon rank test to compare the similarity between the exposed and control users, which was not developed by previous work.

Note that by utilizing the rank test, one is no longer testing that the (weighted) control and exposed groups have the same mean user covariates; rather, it is testing whether the distribution of the user covariates from one group is stochastically larger than the other. Hence the rank test is substantially different than the standard mean test, in terms of both computation and the hypothesis being tested.

We first complete the derivation in Section 3.1 and 3.2, corresponding to two cases: whether there are ties in the data. We then illustrate how the test can be applied to IPW model validation in Section 3.3. The advantage of the proposed rank test over the conventional test are shown in Section 3.4.

3.1 Weighted Mann-Whitney-Wilcoxon Rank Test

The Mann-Whitney-Wilcoxon test statistic is defined as follows: suppose that there are i.i.d. continuous samples $S_1,\dots,S_n$, and i.i.d. samples $T_1,\dots,T_m$, define $U=\sum_{i=1}^n\sum_{j=1}^m \mathbb{I}(S_i \le T_j)$. Under the null hypothesis that S_i's and T_j's are from the same distribution, $u=\frac{U-\mu}{\sigma}$, wtih $\mu=\mathbb{E}[U]=\frac{mn}{2}$ and $\sigma=\sqrt{Var(U)}=\sqrt{\frac{mn(m+n+1)}{12}}$, is asymptotically distributed as $Normal(0,1)$. The Mann-Whitney-Wilcoxon u statistic is an approximation to $\int F(S)dG(T)$, where $S_i \sim F$ and $T_j \sim G$. Most of the generalization of MWW test, e.g., [37], follow the same proof line. We develop a weighted version of the test following similar idea as below.

Suppose we assign a weight to each observation, say we assign $s_1,\dots,s_n$ to $S_1,\dots,S_n$ and $t_1,\dots,t_m$ to $T_1,\dots,T_m$, then $U^*=\sum_{i=1}^n s_i\sum_{j=1}^m t_j\mathbb{I}(S_i \le T_j)$. When there is no tie (i.e. there is not observation such that $S_i = T_j$), we find that $\mu^*=\mathbb{E}[U^*]=\frac{\sum_{i,j}s_it_j}{2}$, and

$$\mathbb{E}[U^{*2}] = \mathbb{E}\begin{bmatrix} & \sum_{i=k,j=l} s_i^2t_j^2\mathbb{I}(S_i \le T_j) \\ + & \sum_{i=k,j\neq l} s_i^2t_jt_l\mathbb{I}(S_i \le T_j)\mathbb{I}(S_i \le T_l) \\ + & \sum_{i\neq k,j=l} s_is_kt_j^2\mathbb{I}(S_i \le T_j)\mathbb{I}(S_k \le T_j) \\ + & \sum_{i\neq k,j\neq l} s_is_kt_jt_l\mathbb{I}(S_i \le T_j)\mathbb{I}(S_k \le T_l) \end{bmatrix}$$
$$= \frac{1}{2}\sum_{i=k,j=l} s_i^2t_j^2 + \frac{1}{3}\sum_{i=k,j\neq l} s_i^2t_jt_l + \frac{1}{3}\sum_{i\neq k,j=l} s_is_kt_j^2$$
$$+ \frac{1}{4}\sum_{i\neq k,j\neq l} s_is_kt_jt_l, \tag{9}$$

which yields

$$\sigma^{*2} = \mathbb{E}[U^{*2}] - \mathbb{E}[U^*]^2$$
$$= \frac{1}{4}\sum_{i=k,j=l} s_i^2t_j^2 + \frac{1}{12}\sum_{i=k,j\neq l} s_i^2t_jt_l + \frac{1}{12}\sum_{i\neq k,j=l} s_is_kt_j^2$$
$$= \frac{1}{12}\left[\sum_{i=k,j=l} s_i^2t_j^2 + \sum_{j,l,i=k} s_i^2t_jt_l + \sum_{i,k,j=l} s_is_kt_j^2\right]. \tag{10}$$

$$\text{Hence } u^* = \frac{U^*-\mu^*}{\sigma^*} \sim Normal(0,1) \tag{11}$$

One can then compare the calculated u^* with the standard normal distribution to test the null hypothesis $H_0 : u^* = 0$ versus alternative hypothesis $H_0 : u^* \neq 0$. If $s_1 = \cdots = s_n = t_1 = \cdots = t_m = 1$, that is, if the samples are equally weighted then $\mu^* = \frac{mn}{2}$ and $\sigma^{*2} = \frac{1}{4}(mn) + \frac{1}{12}nm(m-1) + \frac{1}{12}mn(n-1) = \frac{mn(m+n+1)}{12}$, as expected.

3.2 Weighted Mann-Whitney-Wilcoxon Rank Test with Ties

Now suppose the two samples have ties. Again the test statistic is $u^* = \frac{U^*-\mu^*}{\sigma^*}$. Easily, the estimation μ^* keeps the same. We derive σ^{*2} as follows.

For distinct i, j, and l, obviously

$1 = P(S_i < T_j < T_l) + P(S_i < T_l < T_j) + P(T_j < S_i < T_l) + P(T_j < T_l < S_i) + P(T_l < S_i < T_j) + P(T_l < T_j < S_i) + P(S_i < T_j = T_l) + P(S_i > T_j = T_l) + P(T_j < S_i = T_l) + P(T_j > S_i = T_l) + P(T_l < S_i = T_j) + P(T_l > S_i = T_j)$,

and

$P(S_i < T_j < T_l) = P(S_i < T_l < T_j) = P(T_j < S_i < T_l) = P(T_j < T_l < S_i) = P(T_l < S_i < T_j) = P(T_l < T_j < S_i), P(S_i < T_j = T_l) = P(S_i > T_j = T_l) = P(T_j < S_i = T_l) = P(T_j > S_i = T_l) = P(T_l < S_i = T_j) = P(T_l > S_i =$

$T_j)$.

$$\text{Hence } U^{*2} = \sum_{i=k,j=l} s_i^2 t_j^2 \left[\mathbb{I}(S_i < T_j) + \frac{1}{4}\mathbb{I}(S_i = T_j)\right]$$
$$+ \sum_{i=k,j\neq l} s_i^2 t_j t_l [\mathbb{I}(S_i < T_j)\mathbb{I}(S_i < T_l) + \frac{1}{4}\mathbb{I}(S_i = T_j)\mathbb{I}(S_i = T_l)$$
$$+ \frac{1}{2}\mathbb{I}(S_i < T_j)\mathbb{I}(S_i = T_l) + \frac{1}{2}\mathbb{I}(S_i = T_j)\mathbb{I}(S_i < T_l)]$$
$$+ \sum_{i\neq k,j=l} s_i s_k t_j^2 [\mathbb{I}(S_i < T_j)\mathbb{I}(S_k < T_j) + \frac{1}{4}\mathbb{I}(S_i = T_j)\mathbb{I}(S_k = T_j)$$
$$+ \frac{1}{2}\mathbb{I}(S_i < T_j)\mathbb{I}(S_k = T_j) + + \frac{1}{2}\mathbb{I}(S_i = T_j)\mathbb{I}(S_k < T_j)]$$
$$+ \sum_{i\neq k,j\neq l} s_i s_k t_j t_l \left[\mathbb{I}(S_i < T_j) + \frac{1}{2}\mathbb{I}(S_i = T_j)\right]$$
$$\left[\mathbb{I}(S_k < T_l) + \frac{1}{2}\mathbb{I}(S_k = T_l)\right].$$

$$\text{Thus } \mathbb{E}[U^{*2}] = \frac{1}{2}\sum_{i=k,j=l} s_i^2 t_j^2 + \frac{1}{3}\sum_{i=k,j\neq l} s_i^2 t_j t_l + \frac{1}{3}\sum_{i\neq k,j=l} s_i s_k t_j^2$$
$$+ \frac{1}{4}\sum_{i\neq k,j\neq l} s_i s_k t_j t_l - \frac{1}{4}\sum_{i=k,j=l} s_i^2 t_j^2 P(S_i = T_j)$$
$$- \frac{1}{12}\sum_{i=k,j\neq l} s_i^2 t_j t_l P(S_i = T_j = T_l)$$
$$- \frac{1}{12}\sum_{i\neq k,j=l} s_i s_k t_j^2 P(S_i = S_k = T_j)$$

$$\text{So } \sigma^{*2} = \mathbb{E}[U^{*2}] - \mathbb{E}[U^*]^2$$
$$= \frac{1}{12}\left[\sum_{i=k,j=l} s_i^2 t_j^2 + \sum_{j,l,i=k} s_i^2 t_j t_l + \sum_{i,k,j=l} s_i s_k t_j^2\right]$$
$$- \frac{1}{12}\left[\begin{array}{l} \sum_{i=k,j=l} s_i^2 t_j^2 P(S_i = T_j) \\ -\sum_{i=k,j,l} s_i^2 t_j t_l P(S_i = T_j = T_l) \\ -\sum_{i,k,j=l} s_i s_k t_j^2 P(S_i = S_k = T_j) \end{array}\right]$$

3.3 Weighted Mann-Whitney-Wilcoxon Rank Test with IPW Weights

Again suppose each of the users are assigned weight w_i according to IPW or IPW the TTE estimator. For each of the feature m (indicated by x_{im} for person i), when there is no ties, the test statistic is calculated as $u^* = \frac{U^* - \mu^*}{\sigma^*} \sim Normal(0, 1)$, where

$$U^* = \sum_{i=1}^{N} w_i \sum_{j=1}^{N} w_j \mathbb{I}(x_{im} < x_{jm}) z_i (1 - z_j);$$
$$\mu^* = \mathbb{E}[U^*] = \frac{\sum_{i,j} w_i w_j z_i (1 - z_j}{2};$$
$$\sigma^{*2} = \frac{1}{12}[\sum_{i,j} s_i^2 s_j^2 z_i (1 - z_j) + \sum_{i,j,l} s_i^2 t_j t_l z_i (1 - z_j)(1 - z_l)$$
$$+ \sum_{i,k,j} s_i s_k t_j^2 z_i z_k (1 - z_j)].$$

When there are ties, σ^{*2} is estimated as

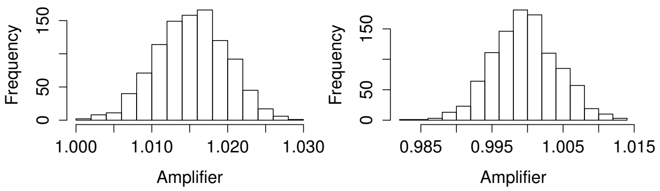
$$\sigma^{*2} = \frac{1}{12}[\sum_{i,j} w_i^2 w_j^2 z_i (1 - z_j) + \sum_{i,j,l} w_i^2 w_j w_l z_i (1 - z_j)(1 - z_l)$$
$$+ \sum_{i,k,j} w_i w_k w_j^2 z_i z_k (1 - z_j)] - \frac{1}{12}[\sum_{i,j} w_i^2 w_j^2 z_i (1 - z_j) P(x_{im} = x_{jm})$$
$$+ \sum_{i,j,l} w_i^2 w_j w_l P(x_{im} = x_{jm} = x_{lm}) z_i (1 - z_j)(1 - z_l)$$
$$+ \sum_{i,k,j} w_i w_k w_j^2 P(x_{im} = x_{km} = x_{jm}) z_i z_k (1 - z_j)],$$

The reduction of the absolute value of test statistic u^* after IPW indicates the balancing effect of the weighting, and the results with real campaigns are shown in Section 4.2.

3.4 Simulation

For a set of 20000 users, we repeatedly generate heavy-tail distributed features with exponential normal distribution. Since the features are generated with continuous distribution, we utilize the rank test with no tie in this simulation. For each of the generated features, we further generate the propensity of exposure and success probability with GBDT, and hence the exposure and success indicators. We assume no causal effect between the exposure indicator and success rates. The simulated datasets are fitted with the proposed method in Section 2, and the covariate balancing are checked with both the standard mean difference method and the proposed weighted rank test.

First we verify that the propensity weighting with GBDT balances the control and exposed groups. We construct the histogram of naive amplifier and the adjusted amplifier in Figure 3 as in Section 2. While the naive estimator is significantly larger than 1, the weighted estimator are centered at 1 with symmetric shape. It is apparent that the weighting successfully captures the biases of the user features.



(a) Naive Estimator

(b) IPW Estimator

Figure 3: Exposure Effect Estimation for Simulated Datasets

In such cases, the weighted features of the control and exposed groups are supposed to be balanced. However as we stated before, the standard mean difference test is vulnerable with heavy-tail distribution. We computed the test statistics of the standard mean difference test for each feature, whose absolute value range from 0.28 to 3.67. Setting the significance level of the hypothesis test to be 0.05 and hence the cut-off value of the test statistics to be 2, 30% of the feature differences are tested to be significantly different than 0. With the proposed rank test, the absolute value of the mean test statistics range from 0.43 to 1.96. Under 0.05 significance level, all of the features pass the rank test. The simulation shows that our rank test is robust when the distribution of user features are heavily skewed, while the conversional test fails to capture the balancing effect of IPW.

We also use the simulated dataset to prove that directly fitting a model with features and exposure indicator is not a sound approach. We fit a logistic regression model and a GBDT model for the simulated success indicator with features and exposure indicators. In the logistic regression model, the coefficient of the exposure indicator has a p-value nearly 0. In the GBDT model, the exposure indicator shows substantial influence. Both of the models mistakenly 'detect' the ad effect on irrelevant conversion, which should not exist.

4. APPLICATION: AD EFFECTIVENESS ASSESSMENT WITH LIVE CAMPAIGNS

In this section we describe three real business use cases, and summarize the results of the three dimensions of ad effectiveness. In all the cases, we collect user-level features, including website visitation, ad exposure, demographic information, market interest, etc., repeatedly on a daily basis. [4] We address the three dimensions of ad effectiveness (uplift, synergy, and audience pool expansion), and show that our framework provides a unified pipeline to produce robust estimator of the ad effectiveness, controlling for variations in user features. We first demonstrate the effectiveness measurement of the campaigns corresponding to uplift, synergy, and audience pool expansion respectively in Section 4.1. The we report the rank test result in Section 4.2.

4.1 Three-Dimensional Assessment of Campaigns

4.1.1 Uplift

We implement the methodology to measure the uplift effect of online ads in two business cases.

The first case analyzes a marketing campaign of a major Internet provider company with only banner ads. We measure the effectiveness of the banner ads comparing to no ad exposure. The success metric is online quotes. The exposed group is defined as the users who were exposed to the banner ads, while the control group subjects were not exposed to ads. The dataset, contains about 18.7 million users, with merely 0.3 million exposed user and 1.9 thousands conversions. This case involves not only sparse successes, but also relatively sparse exposed observations. Hence the subsampling-backscaling strategy includes importance sampling with large sampling rates for the exposed users and converters.

The second case analyzes a marketing campaign of a phone system with only banner ads. We measure the effectiveness of the banner ads comparing to no ad exposure. There are about 0.2M exposed users and 1.2M control users, with 2K converters.

For the Internet provider company campaign, the naive amplifier (as in Equation 4) summarized from the whole dataset is 2.52. With our proposed framework, we reach a population level TTE amplifier 1.751, i.e. the ad lift the conversion rate by 75.1%. The collected amplifier estimations from each chunk have standard deviation 0.137, which suggest small variation in the results for different sub-datasets. The histogram (Figure 4) of the subsample amplifiers shows good robustness of the results. It also shows symmetry and uni-mode, which suggests that the average of the amplifiers from each chunk is a good representation of the amplifier of the population.

For the phone system company campaign, the naive amplifier is 0.51, and the population level TTE amplifier is 1.27. The raw data imply negative uplift effect of the banner ads, while after correcting the biases in the user features of the control and exposed groups, the effect is positive, i.e. the ad lift the conversion rate by 27%. The histogram is similar as the case 2.

In the Internet provider company campaign case, the adjusted amplifier shows dramatic decrease; while in the phone system company campaign case, the adjusted amplifier increases. We utilize these two cases to illustrate our novel interpretation of the result in Section 5.

4.1.2 Synergy

We implement the methodology to measure the joint effect of two advertising strategies on a marketing campaign of a major auto insurance company. The two strategies are a website takeover and a direct response banner. We measure the effectiveness of the website takeover on top of the direct response banner. The exposed group is defined as the users who were exposed to both the website takeover and the banner ads, while the control group subjects were only exposed to the banner ads. The auto insurance company dataset contains approximately 2.8 million users with 11.7 thousand converters.

The naive amplifier is 0.94, and the estimated TTE amplifier is 1.184, i.e. the webpage takeover lift the conversion rate by 18.4% on top of the direct response banner ad. The result is shown in the histogram Figure 5. This shows that naive amplifier underestimates the amplification effect of the two advertising strategies, but in fact, users who were exposed to both strategies are 1.184 times more likely to convert.

4.1.3 Audience Pool Expansion

We again consider the marketing campaign of the auto insurance company as in Section 4.1.2, but measure the reach extension effect of the upper-funnel placement (website takeover) on the lower-funnel placement (direct response). We measure how much more likely the users migrate into interest segments that can then be targeted by the direct response campaign after exposed to website takeover campaign. The success metric is the indicator representing whether or not each user is included in the targeting pool of the lower-funnel placement. The exposure is defined as exposure to the upper-funnel ad impressions. The naive amplifier is 1.80, and the estimated TTE amplifier is 1.23. Thus, the webpage takeover brings 23% more audience to the direct response banner ad. The result is shown in the histogram Figure 6.

4.2 Check the Balancing Effect of Propensity-Based Weighting

The check the balancing effect of the weighting, we implement our novel rank test and the result shows significant reduction of test statistics after weighting. The percentage reduction ranges from 73.0% to 92.3%. The result verifies that the weighting significantly balanced the relevant user features. [5]

To give an visualized example of the balancing effect of the IPW on user characteristics, we summarize the total network activity before and after the weighting in the Internet provider campaign in Figure 7. The figure shows significant improvement in the balance of network activity. Similar phenomena are observed for other important features, such as auto purchase intention in the auto insurance dataset. This is consistent with the rank test results.

[4] The reported dataset and results are deliberately incomplete and subject to anonymization, and thus do not necessarily reflect the real portfolio at any particular time.

[5] The conventional standardized mean test show the percentage reduction ranges from 50.0% to 77.6%. However we have shown that the proposed rank test yields more accurate results with advertising data, where the user characteristics are usually skewed and heavy-tailed.

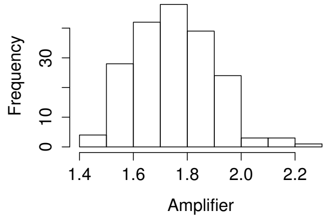


Figure 4: TTE Amplifiers, Uplift

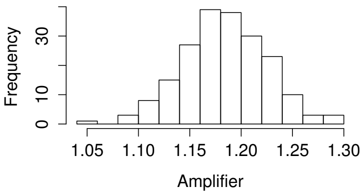


Figure 5: Synergy

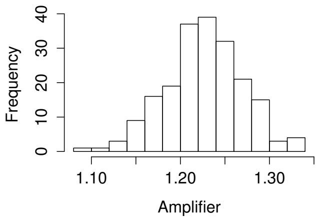
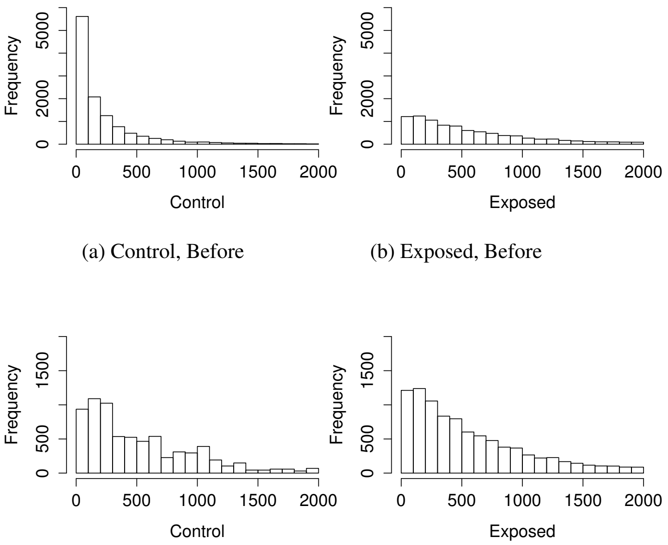


Figure 6: Audience Pool Expansion



(a) Control, Before

(b) Exposed, Before

(c) Control, After

(d) Exposed, After

Figure 7: Network Activities Before (a, b for control and exposed groups respectively) and After (c, d for control and exposed groups respectively) the Weighting

5. INTERPRETATION: UNDERSTANDING THE ADJUSTED AMPLIFIER

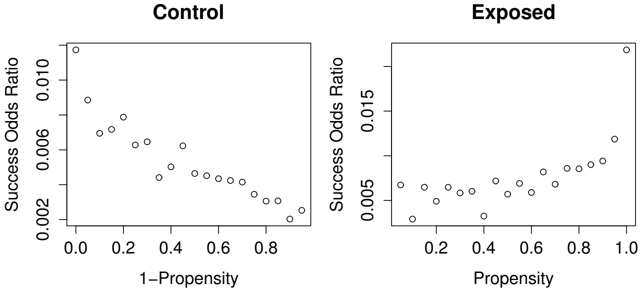
In all the marketing campaigns in Section 4, the analysis reveals positive ad impact on uplift, synergy, and audience pool expansion aspect. However, the change of amplifier after causal inference can be positive or negative, which requires further interpretation from business point of view. In this section we compare of the raw amplifier and adjusted amplifier after causal inference. Note that the causal inference model 'corrects' the amplifier by eliminating the effect of user features, and hence the change of the amplifier reveals the nature of the ad placement: either it is doing 'smart cheating' and reaching users who would convert even without the ad, or reaching users who would not convert without the ad. There are two possible scenarios:

1) The first scenario is that the amplifier decreases after adjustment. This means the confounding effect of user features inflates the raw amplifier, and hence the exposed group is doing 'smart cheating', namely, the exposed group contains more users who are likely to convert even without ad exposure.

In the Internet provider company campaign case in Section 4.1.1, the amplifier shrinks after adjustment. To further investigate the users in the control and exposed groups, we calculate the success odds ratio of both groups along with the probability belonging to the corresponding group, as in Figure 8. The increasing trend in Figure 8(b) shows that the exposed group tends to contain users who are more likely to convert, and the control group the opposite. Hence the placement is doing 'smart cheating', and the causal inference eliminates such effect by shrinking the amplifier, as expected.

2) The second scenario is that the amplifier is enlarged after adjustment. This means that the confounding effect of user features deflates the raw amplifier, and hence the exposed group is reaching 'hard users', namely, the exposed group contains more users who are less likely to convert without ad exposure.

In the phone system company campaign case in Section 4.1.1, the amplifier is about twice after adjustment. We draw the success odds ratio of both groups along with the probability belonging to the corresponding group as in Figure 9. The declining trend in Figure 9(b) shows that the exposed group tends to include users who are less likely to convert, i.e. 'hard users', and the control group has more 'easy users'. Hence the causal inference eliminates such effect, and brings back the true impact of ads.



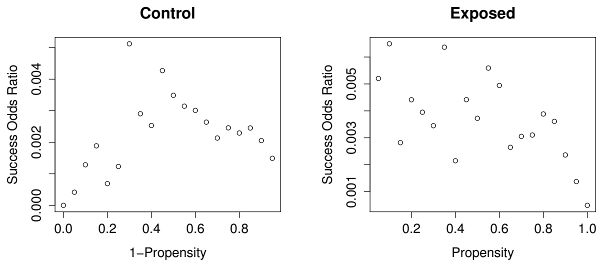
(a) Control

(b) Exposed

Figure 8: Success Odds along with Probability Belonging to Corresponding Group, Internet Provider

6. CONCLUSIONS

In this paper, we construct a unified framework to measure ad effectiveness from observational data. Our solution incorporates IPW estimator for the causal inference? a novel robust rank test for model validation. We also investigate the change of amplifier before and after IPW adjustment, to find the 'smart cheating' in ads. In the multi-treatment framework, the validation and smart cheating detection approaches proposed in this paper still holds. One can calculate the success odds ratio of each group along with the probability belonging to the corresponding group, and find out the ad strategy that tends to include user who are more like to convert,



(a) Control (b) Exposed

Figure 9: Success Odds along with Probability Belonging to Corresponding Group, Phone System

i.e., doing 'smarting cheating'. The validation method in Section 3 may also be generalized accordingly, to conduct simultaneous comparison of the features of users receiving different treatments.

The framework provides a thorough solution to the ad effectiveness measurement, including uplift, synergy, and audience pool expansion effect, which is crucial for online advertising. Also, this paper focuses on measuring the effectiveness of online ads, but the framework is readily applicable to measure the effectiveness of other kinds of treatments on various user metrics, for example the impact of different strategies on user engagement metrics. It may serve as a supplement or substitute to experiments, in the areas such as recommender systems and web search, where controlled experiments are extensively used for comparing algorithms and models.

7. REFERENCES

[1] D. Agarwal, L. Li, and A. J. Smola. Linear-time estimators for propensity scores. In *International Conference on Artificial Intelligence and Statistics*, pages 93–100, 2011.

[2] J. Barajas, J. Kwon, R. Akella, A. Flores, M. Holtan, and V. Andrei. Marketing campaign evaluation in targeted display advertising. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, page 5. ACM, 2012.

[3] A. Basu, D. Polsky, and W. G. Manning. Use of propensity scores in non-linear response models: the case for health care expenditures. Technical report, National Bureau of Economic Research, 2008.

[4] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] D. Chan, R. Ge, O. Gershony, T. Hesterberg, and D. Lambert. Evaluating online ad campaigns in a pipeline: causal models at scale. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–16. ACM, 2010.

[6] Y. Chang and E. Thorson. Television and web advertising synergies. *Journal of Advertising*, 33(2):75–84, 2004.

[7] D. R. Cox. Planning of experiments. 1958.

[8] B. Dalessandro, C. Perlich, O. Stitelman, and F. Provost. Causally motivated attribution for online advertising. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, page 7. ACM, 2012.

[9] A. Dasgupta, K. Punera, J. M. Rao, X. Wang, J. Rao, and X.-J. Wang. Impact of spam exposure on user engagement. In *USENIX Security*, 2012.

[10] G. M. Fulgoni and M. P. Morn. Whither the click? how online advertising works. *Journal of Advertising Research*, 49(2):134, 2009.

[11] S. Guo and M. W. Fraser. Propensity score analysis. *Statistical methods and applications*, 2010.

[12] J. J. Heckman, H. Ichimura, and P. Todd. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.

[13] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[14] K. Imai. Do get-out-the-vote calls reduce turnout? the importance of statistical methods for field experiments. *American Political Science Review*, 99(2):283–300, 2005.

[15] K. Imai, G. King, and E. A. Stuart. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society)*, 171(2):481–502, 2008.

[16] M. Lechner. Earnings and employment effects of continuous gff-the-job training in east germany after unification. *Journal of Business & Economic Statistics*, 17(1):74–90, 1999.

[17] S. F. Lehrer and G. Kordas. Matching using semiparametric propensity scores. *Empirical Economics*, 44(1):13–45, 2013.

[18] R. A. Lewis, J. M. Rao, and D. H. Reiley. Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of the 20th international conference on World wide web*, pages 157–166. ACM, 2011.

[19] R. T. Lindsay, T. Carriero, and Y.-F. Juan. Measuring impact of online advertising campaigns, May 26 2009. US Patent App. 12/472,318.

[20] B. Lu, E. Zanutto, R. Hornik, and P. R. Rosenbaum. Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, 96(456):1245–1253, 2001.

[21] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 18(1):50–60, 1947.

[22] D. F. McCaffrey, G. Ridgeway, A. R. Morral, et al. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403–425, 2004.

[23] S. P. Novak, S. F. Reardon, S. W. Raudenbush, and S. L. Buka. Retail tobacco outlet density and youth cigarette smoking: a propensity-modeling approach. *American Journal of Public Health*, 96(4):670–676, 2006.

[24] E. L. Olson and H. M. Thjømøe. Sponsorship effect metric: assessing the financial value of sponsoring by comparisons to television advertising. *Journal of the Academy of Marketing Science*, 37(4):504–515, 2009.

[25] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

[26] P. R. Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032, 1989.

[27] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[28] P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

[29] P. R. Rosenbaum and D. B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

[30] D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[31] Y.-B. Song. Proof that online advertising works. *Atlas Institute, Seattle, WA*, 2001.

[32] O. Stitelman, B. Dalessandro, C. Perlich, and F. Provost. Estimating the effect of online display advertising on browser conversion. *Data Mining and Audience Intelligence for Advertising (ADKDD 2011)*, 8, 2011.

[33] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[34] B. Villalonga. Does diversification cause the" diversification discount"? *Financial Management*, pages 5–27, 2004.

[35] P. Wang, Y. Liu, M. Meytlis, H.-Y. Tsao, J. Yang, and P. Huang. An efficient framework for online advertising effectiveness measurement and comparison. *WSDM 2014 proceedings*, 2013.

[36] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[37] J. Xie and C. E. Priebe. Generalizing the mann-whitney-wilcoxon statistic. *Journal of nonparametric statistics*, 12(5):661–682, 2000.