# The Invisible Barrier: The Effect of Promoting Agencies on Sales in Electronic Markets for Music


Marios Kokkodis
Boston College
kokkodis@bc.edu

Theodoros Lappas
Stevens Institute of Technology
tlappas@stevens.edu

Konstantinos Pelechrinis
University of Pittsburgh
kpele@pitt.edu

## Introduction

No barrier-to-entry electronic marketplaces have been growing in the past decade in a parallel trajectory with the e-commerce explosion. For instance, Amazon books allows independent writers to publish their work, Apple's AppStore gives the option to ambitious developers to merchandise their apps, and Beatport.com facilitates music producers with a platform to sell their songs. All these marketplaces have experienced a significant growth in recent years; For example the AppStore has reached a revenue of $10 billion in 2014, twice as much as the previous year[1].

The operational premise that underlies these marketplaces is simple and promotes both equality and diversity: any individual artist with the ability to create an original book/ app/ song, is eligible to make it available on these platforms and sell it for profit. The marketplaces make their revenue by taking a commission on the sales – the AppStore for example takes 30% off.

No barrier-to-entry markets can be thought of as a natural part of the evolution of the way goods are distributed in the Internet era. For example, the music industry, which is the focal point of this work, has evolved tremendously in the past 20 years (Koster 2011); From record/tape/cd sales to illegal P2P sharing, to the emergence of iTunes and finally to the battlefield of streaming services such as Pandora, Spotify and Apple music. In parallel, music has been digitally transformed: from the 60's analog classic rock, to the 70's heavy metal, to 80's disco, 90's Hip Hop/alternative rock to the millennials completely digital electronic scene. This complete digitalization of music in the way it is being distributed, produced and consumed, facilitated the increased market shares of independent music labels[2], which in turn empowered the emergence of new no barrier-to-entry music marketplaces such as `Beatport.com`.

Existing literature has established a link between product promoting/advertising and product sales (Assmus et al. 1984, Clarke 1976, Blattberg and Jeuland 1981). Specifically to the music industry, it takes only a look on Billboard.com[3] to verify that music labels – which are the main promoting agencies – are occupying the bulk of the charts. This suggests that even though marketplaces such as Beatport.com facilitate music producers with a platform to publish their productions, independent produces, who are not being promoted by a major music label, experience **invisible barriers** on the road to success.

Driven by these observations, in this work, we are interested in examining whether no barrier-to-entry markets for music are in practice holding on to their premise of fair competition. In particular, we focus

[1] http://goo.gl/TBC2qk
[2] http://goo.gl/iAIH3n
[3] http://www.billboard.com/charts/hot-100

on electronic markets for music, and central to our study is the question of **what is the effect of music labels on a song's propensity to enter the charts in a zero barrier-to-entry electronic market**. To address this question we collect and analyze a unique panel dataset from a major marketplace of electronic music, `Beatport.com`. From the collected data, the only instances that have deterministic outcome are the songs that have already entered charts (i.e.,, positive only instances). However, to answer our main question we need both positive and negative instances. To overcome this, we employ a survival analysis to make stochastic estimates about a song's likelihood to enter the charts. The result of this analysis is a dataset that includes both positive and negative instances. We next use this observational dataset to build a multidimensional "tree-based causal inference" framework to study how the interaction of different music label characteristics affect the placement of a song in the charts. Our findings suggest that certain interplays among the features of promoting agencies (i.e., music labels) increase the probability of success up to nine times.

Our work contributes to the current literature in promotion, sales and the music industry by (1) identifying a strong effect of music labels on sales in electronic marketplaces for music and (2) by attributing this effect to multiple dimensions of the treatment. Furthermore, we present a unique and widely applicable methodological framework that combines parametric survival analysis and a cutting-edge quasi experimental causal inference technique. From a managerial perspective, our work reveals the problematic rich-get-richer phenomenon that prevails on Beatport.com. To alleviate this, we propose two actionable solutions: (1) control for the prevalence of music labels in the current ranks and (2) create separate ranks for independent artists.

## Background

Our goal in this work is to study whether no-barrier to entry markets hold up to their promise, by creating an equal opportunity environment for all participants. Even though we are the first to study this particular question, previous related works have dealt with (1) the effect of advertising and promotion on sales and (2) various aspects of the music industry. In this section we connect these previous studies to our work, and we conclude by summarizing our contributions.

### Promotion, Sales and the Invisible Barrier

The focal point of this work is to quantify the effect of music labels on sales in a particular electronic marketplace for music. Previous works in marketing research have already established a clear strong link between advertising and sales (Assmus et al. 1984, Clarke 1976, Blattberg and Jeuland 1981). In our setting,

the bulk of advertisement takes place in social media, either through promoted posts[4], or through personal posts by the artist's and/or music label's fan-page. We know already from previous work that broadcasting in social media has a significant effect on sales (Chen et al. 2015). Hence, independent and new artists face a disadvantage: they neither have the budget to run promotional campaigns nor the critical mass of followers to broadcast their new music productions.

This barrier becomes even bigger when we consider other dimensions of these marketplaces. Diversified recommender systems could have pose a viable solution to this problem, but standard online recommender systems (e.g., collaborative filtering) that are commonly deployed by online marketplaces have been found to magnify the rich-get-richer effect (Fleder and Hosanagar 2009). Word of mouth could have also been a solution – since we know that it is as effective as promotional campaigns (Lu et al. 2013) – however in our scenario, mainstream reviewing outlets for music producers are practically non-existent. Finally, independent artists have to fight against both the rank positioning effects (Ghose and Yang 2009, Animesh et al. 2011, Ghose et al. 2014) and the social influence effects (Salganik et al. 2006) which further empower the rich-get-richer phenomenon (despite the fact that these effects are not tied to the actual product quality (Salganik et al. 2006)).

For all these reasons, we expect that no-barrier to entry electronic markets are failing to hold on to their expected behavior. In this work we aim to quantify the effect that music labels have on sales, and propose ways to minimize the spread between the "rich" and the "poor", i.e., artists represented by well established music labels vs. independent producers.

## The music industry

Music is an experience good (Nelson 1970), that is, the true quality of a song is subjective, and it is only revealed when the consumer listens – experiences – a particular song. As an experience good, music has always being tremendously affected by current trends and promotions. Towards this direction, previous works studied the interplay between (1) blogging and music sampling (Dewan and Ramaprasad 2012), (2) radio play and future sales (Dewan and Ramaprasad 2014) and (3) blog posts and future sales (Dhar and Chang 2009). Our study is orthogonal to these works since (1) we focus on a completely different marketplace with very distinct characteristics and appeal to independent artists, and (2) because our main question examines the effect of music labels on individual song sales–and not on the traditional offline album sales.

A different line of research on the topic focuses on estimating the survival time of a song in the charts. A series of features have been found in the past to affect this survivor probability of a song, including (1) the type

[4]For example "sponsored" posts on Facebook, https://goo.gl/jCc2q2

of album, the seasonal demand and the initial popularity (Strobl and Tucker 2000) and (2) the introduction of peer to peer song sharing (Bhattacharjee et al. 2007). Taking this analysis one step further, researchers have also proposed models that predict the simultaneous movement of multiple items up and down the charts over time (Bradlow and Fader 2001). Contrary to these works that focus on understanding the survival evolution of a song in the ranks, we are interested in identifying factors (i.e., music label characteristics) that increase (or decrease) the likelihood of a particular song entering the charts.

Finally, other related research has dealt with estimating the sales of a new album (Lee et al. 2003) but also with identifying optimal strategies and pricing for digital music (Danaher et al. 2014). These two dimensions are not closely related to this piece of work since our objective is neither sales prediction nor optimal marketing/pricing.

### Reputation in Electronic Markets

Finally, as we mentioned earlier, reputation systems have the potential of balancing out some of the promotional effect of music labels. However, at this point, such mechanisms are not present in electronic markets from music. Nevertheless, we know from previous works that reputation mechanisms have a multidimensional causal impact on online marketplaces, from resolving information asymmetries (Dellarocas 2006, Kokkodis and Ipeirotis 2015) to improving transaction efficacy (Bakos and Dellarocas 2011, Bolton et al. 2004), and to affecting sales and willingness to pay (Chevalier and Mayzlin 2006, Resnick et al. 2006) It is clear that these studies have different focal points from our work – they are mentioned here for completeness.

### Contribution of our work

Our work extends the current literature in promotion, sales and the music industry by contributing mainly in four dimensions. First, we present a novel study that identifies the effect of music labels on sales in no-barrier-to-entry electronic marketplaces. Second, we are able to quantify this effect by splitting it into multiple dimensions of the music labels. Third, based on our results, we are able to propose a series of actions that this particular electronic marketplace could employ in order to alleviate the "music label" effect. Finally, from a methodological perspective, we present a novel and technically sound approach for estimating the probability of a song entering the charts, as well as we employ a tree-based causal inference approach that allow us to study the multidimensional effect of music labels from observational data.

## Data and Experimental Setup

The digitalization of music in the way it is being distributed as well as the way it is being produced facilitated the emergence of new electronic marketplaces like `Beatport.com`. In this section we describe in detail the characteristics of this electronic music marketplace we are studying, along with the dataset we collected and used for our study. We further present the set of covariates we consider in our analysis. In the rest of this paper, the "big three" record labels represent Sony, Warner and Universal music.

### The Electronic Market of Beatport

To employ our proposed framework (discussed in detail in the next section) we use a unique dataset from `Beatport.com`. Beatport is the major outlet for electronic music. As a private company, Beatport does not report statistics about its growth and revenue. However it was revealed in 2012 that it featured more than 200,000 register DJs (including "superstars" such as Tiesto, Avicii and DeadMau5), while it attracted around 40 million unique visitors[5]. Online users have the ability to visit Beatport, discover the latest music productions, and decide whether or not to purchase songs.

Beatport features songs across a total of 23 genres; For each one of these genres, Beatport creates a different ranking. These rankings are proportional to the sales of the song. A screenshot of the top 100 ranked songs of the genre "House" is shown in Figure 1. We see the name of each song, along with the artist(s) and the music label. We further notice that the marketplace allows users to listen to a sample of the song – bottom part of the figure – before they purchase it. The music labels that participate on Beatport are mostly independent labels (such as Andjunabeats, Ultra and Spinnin' records – shown in Figure 1). Nevertheless, we also find releases from labels that are subsidiaries of the "big three" record companies (e.g., Jive records).

Beatport rankings can have a defining impact on the producers' market shares in a booming global industry of $6.9 billion[6]. Producers that appear repeatedly on the charts, become popular, gain exposure, create a big fan base, and become able to headline large music festivals with hundreds of thousands of attendees. This exposure creates a positive feedback loop, which in turn is being monetized by the artists prolonged stay in the charts. As an illustrative example, let us consider Martin Garrix. Three years ago, Martin Garrix was an unknown to the public teenager/DJ. In 2013, the 16-year old Garrix produced the song "Animals", which reached No. 1 on Beatport and stayed there for 27 days. After this huge success, Garrix became one of the most wanted DJs to

[5] http://goo.gl/4iv9ni
[6] http://goo.gl/IBaKGQ

headline festivals around the world, reaching in 2015 an estimated net worth of $14[7] million.

## Dataset

Our dataset consists of songs that have been released on Beatport between March $7^{th}$ and May $11^{th}$ 2015. In particular, we collected a total of 34,949 songs, produced by 8,298 artists and promoted by 3,461 independent labels[8]. Out of these songs, only 2,043 (5.8%) appeared to at least one of the 23 top-100 genre-specific rankings. With $D$ denoting the random variable that describes the number of days it takes for a song to appear for **first time** in the rankings, Figure 2 depicts the per-genre probability density of $D$. On average (across genres), it takes 11 to 12 days for a song to appear in the ranks after its release. Furthermore, the probability density for $D > 40$ is extremely small.

## Covariates

As alluded to above our goal is to study the effect of promoting agencies (i.e., music labels) on the probability of a song to appear in the ranks. Since, we are dealing mostly with independent labels, there is not a public metric to use that will rank these labels based on their market share and/or impact. In order to capture the relative market share of each independent label $i$ that appears in our dataset we first compute the number of releases ($rl$), which intuitively represents the size of the label. All else being equal we would expect that a music label with a higher number of releases will have a higher market share and/or impact on the market. Furthermore, the spread of the label releases across the different genres might also be a representative factor that the label's prevelance. To capture this spread we estimate the entropy of each label across the available genres Finally, intuition suggests that the trending of a label (denoted with $tl$) is descriptive of the label's (current) market value. To capture this trending we count the total number of songs that a label had in the charts the week before the release date of a song that is under consideration. Note that in order to estimate these three quantities for all the 3,461 independent labels we analyze a total of 566,720 songs released between 2000 and 2015.

Furthermore, in order to control for confounding factors that might have affect the placement of a song in the top ranks, we compute a set of song/artist-specific characteristics. In particular, we first estimate the popularity of the artist who produced the song; To do so, we collect the total number of Twitter followers for each one of the 8,298 artists in our dataset the ($f$). Furthermore, we compute the number of total releases

[7] http://goo.gl/jJOjlO

[8] The dataset and the python code used for this study can be made available upon request.

of each one of those artists, ($ra$), and finally, we estimate the entropy of each artist ($ea$) across the available genres.

### Dataset Statistics

In Table 1 we present the statistics of all the six variables we consider in our analysis. Furthermore, in Table 2 we show a representative set of artists in the top, middle and bottom tier – i.e., $33^{th}$ percentile – of our dataset. We observe a significant diversity with respect to the artist names ranging from popular "superstars" (e.g., Britney Spears, Bruno Mars and will.i.am), to very successful producers that populate festivals around the world (e.g., Myon & Shane 54 and Hot Since 82) to completely unknown to the general public producers – bottom tier. Finally, Table 3 presents examples of top, middle and bottom tier labels in the Beatport marketplace. Note that all of these labels are independent (i.e., they do not belong to one of the "big three" record labels).

Thus far we have described in detail the dataset that we will use in order pursue our study. Next, we focus on the first part of our analysis, and in particular on how to predict whether or not a song will appear in the charts.

## A song's survival model

In this section we draw on statistics literature and propose a survival analysis to estimate the conditional probability that a song will enter the charts given a set of confounding factors. We start by arguing why this approach is appropriate, followed by the mathematical modeling. Finally, we discuss the results from the analysis of our data and their usefulness for the next part of our study.

### Accelerated Failure Time Models

For the purpose of identifying the effect of a music label on the placement or not of a song in the charts we need to have both negative – songs that did not make it in the charts – and positive – songs that made it in the charts – instances. In our setting, there is always a non-zero probability for any song to appear in the charts. However, as we have presented in Figure 2, there is a vanishingly small probability for this happening 40 days after the release date of the song. A straightforward, but rather naive approach would be to consider as negative instances all the songs that did not make it in the charts within a window of days after its release. This approach is problematic for three major reasons: **(i)** it constrains our analysis on instances that are at least of a specific age. This is misleading because such an arrangement leads to a comparison of positive and

negative instances that potentially appeared in completely different time periods (we discuss this in more detail in the next section). **(ii)** This approach implicitly assigns equal probabilities to all instances based on only one variable, that is, time. It explicitly ignores the rest of the observed characteristics of each instance and focuses only on time. Intuitively we believe that a set of other factors might be correlated with the desired probability, and as so, we would like to use this information as well. Finally, **(iii)** by employing such an ad-hoc approach there is no formal way to choose the age threshold, which can have significant implications on the final results.

To avoid these pitfalls, we propose a more systematic approach that draws on the survival modeling literature. Survival models associate the time before some event occurs to a set of covariates that might be correlated with both the event and the time lapsed. These models are typically used to answer questions such as "which portion of a given population will survive past a certain time?". Because of the nature of the problems that survival analysis has been applied to, the actual occurrence of the expected event is referred to as *death*. In our setting, the *death* corresponds to the first time that a song makes it in the charts after its release.

In general, with $T$ being the random variable that describes the *death time*, the cumulative survival probability of a song at time $t$ can be described by $S(t) = \Pr(T \geq t) = 1 - \Pr(T < t)$. In our case, we are interested in estimating the conditional cumulative survival probability of an instance $i$, $S(t|\boldsymbol{X}_i)$, where $\boldsymbol{X}_i$ is the vector of covariates described earlier in the previous section.

The most commonly used survival model in the literature is the Cox model of proportional hazards (Cox and Oakes 1984). The Cox model assumes that the effect of a unit increase in a covariate is multiplicative with respect to the hazard function[9]. By testing for this proportionally (Grambsch and Therneau 1994) we found that our set of variables violates this assumption, suggesting that Cox model is not appropriate for our setting.

A different set of survival models fall under the Accelerated Failure Time (AFT) cluster. These are parametric models, where the survival probability is assumed to follow a given distribution. Our preliminary analysis presented in Figure 2 suggests that we can employ an AFT model with log-normal distribution for the survival probability in our setting. The log-normal distribution assumes that initially the probability of dying, that is, appearing in the top-100 ranks, increases up to a point and then starts decreasing with time. Our data exhibits similar behavior, since the probability of a song entering the charts increases for approximately the first week and then it starts decreasing.

In general, an AFT model describes the death time $t_i$ of song $i$ as $\log(t_i) = \boldsymbol{\beta}\boldsymbol{X}_i + \varepsilon$ . (Cleves 2008). The word "accelerated" describes the fact that an underlying distribution is assumed for $\tau_i = \exp(-\boldsymbol{\beta}\boldsymbol{X}_i)t_i$ ,

[9]The hazard function is defined as $\lambda(t) = \dfrac{S'(t)}{S(t)}$ (Cleves 2008).

where $\exp(-\boldsymbol{\beta}\boldsymbol{X}_i)$ is called the accelerating factor. If this factor is positive then time moves faster – so failure is expected to occur sooner – and vice versa.

In the log-normal case, we assume that $\tau$ follows a log-normal distribution, i.e., $\tau_i \sim \log\mathcal{N}(\beta_0, \sigma^2)$. Hence, taking the log on both sides of the previous Equation we get $\log(t_i) = \boldsymbol{\beta}\boldsymbol{X}_i + \log(\tau_i)$. Now since $\tau_i \sim \log\mathcal{N}(\beta_0, \sigma^2)$ then $\log(\tau_i) \sim \mathcal{N}(\beta_0, \sigma^2)$. Hence, we can further re-write the previous equation as $\log(t_i) = \beta_0 + \boldsymbol{\beta}\boldsymbol{X}_i + u_i$, which converts the problem into a linear regression problem where the error $u_i$ is distributed normally with mean 0 and standard deviation $\sigma$. As a result we get $E[\log(t_i|\boldsymbol{X}_i)] = \beta_0 + \boldsymbol{\beta}\boldsymbol{X}_i$ .

Since an AFT model has the characteristic of accelerating the event by a constant factor over a baseline model, we can estimate this baseline model by zeroing the variables of vector $\boldsymbol{\beta}$. The baseline survivor function of $t$ now becomes $S_0(t) = 1 - \Phi\left(\frac{\log t - \beta_0}{\sigma}\right)$ , where $\Phi$ is the cumulative normal distribution. The last step is to compute the conditional survival cumulative distribution, $S(t|\boldsymbol{X}_i)$. We know that the proposed AFT model will accelerate the previous baseline by a factor of $\exp(-\boldsymbol{\beta}\boldsymbol{X}_i)$. Hence we get:

$$\begin{aligned} S(t_i|\boldsymbol{X}_i) &= S_0\big((\exp(-\boldsymbol{\beta}\boldsymbol{X}_i)t_i\big) \\ &= 1 - \Phi\left(\frac{\log\big(\exp(-\boldsymbol{\beta}\boldsymbol{X}_i)t_i\big) - \beta_0}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{\log(t_i) - (\beta_0 + \boldsymbol{\beta}\boldsymbol{X}_i)}{\sigma}\right) \end{aligned}$$

Now we are ready to fit this model in our data to create the desired set of negative instances.

### Obtaining Negative Instances

In our analysis, the vector of covariates is defined as $\boldsymbol{X}^T = [tl, \log(rl), el, \log(ra), ea, \log(f)]$. In order to estimate the coefficients $\boldsymbol{\beta}$ we fit a model on a non-censored dataset. The latter includes all the instances that have *died* within the examined window, i.e., the 2,043 positive instances we have collected. The results are shown in Table 4. First, we observe that our model is statistical significant at the significance level of $\alpha = 0.001$. Next, the covariates $tl, \log(rl), \log(f)$ and $el$ are statistically significant also at the significance level of $\alpha = 0.001$. All the positive coefficients prolong the death time of a song. For example, a unit increase of $\log(rl)$ will decelerate the appearance of a song in the charts by a factor of $\exp(0.094) = 1.1$ (i.e., will decelerate by $\sim 10\%$). On the contrary, the negative coefficients expedite the entrance of a song in the charts. Hence, all else being equal, a unit increase of $tl$ will accelerate the dying process by a factor of $\exp(-0.025)) = 0.97$ (i.e., will accelerate by 3%). This indicates that higher trend $t_l$, higher number of twitter followers $f$ and higher number of label releases $rl$ are associated with faster placement in the charts. Note that

these effects are under-estimated, since we do not include in our survival analysis the right-censored instances that have not entered the charts yet. This does not undermine the significance of our modeling approach, since are ultimate goal is to identify negative instances.

Having learned this model on the positive instances, we apply it on the rest of 31,488 songs that have not yet appeared on the charts. We then use the predicted probabilities to create a set of songs that with very high probability should have already died yet they stand alive. It is then rational to assume that this set of songs will never get in the charts, and hence, assign them a negative label. Formally, we assign a negative label to the set of instances for which $t_i^* < d_i$ , where $t_i^*$ is the predicted day for which the survival probability for song $i$ becomes vanishingly small, i.e., $S(t_i^*|\boldsymbol{X}_i) < \eta$ (we choose $\eta = 0.005$), and $d_i$ is the number of days after the release of song $i$. The resulting set includes 15,377 negative instances.

To summarize, the survival analysis proposed here alleviates all three problems mentioned in the beginning of this section. First, by employing the AFT model we create a dataset with both positive and negative instances with songs that have been released in the same period – a feature that is important to control for unobserved external shocks (discussed over the next section). Second, the AFT approach assigns more realistic probabilities of success based on the actual values of the variables in $\boldsymbol{X}$. Finally, our methodology is formal and avoids any ad-hoc selection of thresholds/parameters, besides the parameter $\eta$ which represents a very small probability[10]. At this point, we have the dataset that we can use to estimate the effect of promoting agencies on the success of a track.

## The Effect of Music Labels

In this section we use the resulting set of positive and negative instances to study the effect of music labels on the placement (or not) of a song in the charts. We start by describing the methodology we used, and then we present our findings.

### Tree-based Causal Inference

Our goal is to examine the presence of a causal relationship between characteristics of the music label of a song with the song's success. This is rather challenging given the fact that we have access to observational data. Hence, we need to utilize quasi-experimental techniques (Shadish et al. 2002). These techniques are very powerful but in their majority deal with scenarios where only a single treatment is present. In our setting, the

[10]Note here that, if we had a small set of true negative instances, parameter $\eta$ could also be formally set based on the false positive rate that we can tolerate.

treatment effect comes from the presence of music labels on Beatport.com, which as we mentioned earlier have a set of characteristics (see Tables 1 and 3). This multidimensionality of music labels suggests that whenever a given label releases a new song, that particular song is *treated* based on the interactions of certain label characteristics. Hence, we need a methodology that allow us to capture a multidimensional treatment effect.

Wang *et al.* Wang et al. (2015) proposed a new, tree-based causal inference methodology that efficiently considers multidimensional (categorical or continuous) treatments. Naturally, this approach perfectly fits the purpose of our study. In particular, the tree-based causal inference algorithm proceeds as follows: we first build decision trees that estimate the probability of an instance to be treated given the feature vector $\boldsymbol{X}_i$, $\Pr(\boldsymbol{Z}_i|\boldsymbol{X}_i)$, – $\boldsymbol{Z}_i$ is a vector that represents the multidimensional treatment of instance $i$. The tree structure guarantees that the conditional distribution of treatment becomes homogenous within each leaf. Next, the treatment effect within each leaf node is estimated, followed by an estimation of the weighted average across all the leaf nodes. The complete process is presented in Algorithm 1.

To reiterate, the treatment in our case is multi-dimensional. In particular, we consider three variables that describe the characteristics of a the music label: $\log(rl),\ el$ and $tl$ (see Table 1). In order to overcome the natural sparsity that a tree-based classification can cause within the leaves when continuous treatments are considered, we quantize all the three treatment variables. Essentially, we transform the treatment vector $\boldsymbol{Z}_i$ into a binary vector, where all possible treatments form a set $\mathcal{Z}$ where $|\mathcal{Z}| = 8$. In particular for each treatment variable $z \in \{\log(rl),\ el,\ tl\}$ we define the following:

$$\mathrm{Z}_{i,z} = \begin{cases} 1, & \text{if } z(i) > Quartile_1(z) \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where $z(i)$ is the value of treatment $z$ for instance $i$, and $Quartile_1(z)$ is the median of the values of $z$ in our dataset.

Note here that while we could have used other quantiles, we chose to use the median since this provides a close to *balanced* split of our instances to the 8 different treatments. Other threshold choices lead to treatments that are not represented in our dataset. We elaborate on this in our last section. In Table 5 we show all the treatment combinations along with the number of positive and negative instances. The last step is to build the decision tree that estimates the conditional probability of an instance to be treated, $\Pr(\boldsymbol{Z}_i = \boldsymbol{z}|\boldsymbol{X}_i)$, where $\boldsymbol{z} \in \mathcal{Z}$. The feature vector $\boldsymbol{X}$ in this case is different from the vector we have used earlier in Section . First, we do not consider the three label-related variables that are defining the treatment set. Second, we include a

new variable, $dr$, which represents the number of days between the release of the song and the starting date of our data collection process, March $7^{th}$ 2015. This variable is important because it allows us to control for unobserved exogenous shocks that might have affected the whole industry. The final feature vector hence becomes $\boldsymbol{X}_i = [\log(ra), ea, \log(f), dr]$.

## Results

In Figure 3 we show an example of a pruned tree for our case – the actual tree is much larger, with a total of 89 leaves. The darker nodes represent decision points (left 'True', right 'False'). The lighter shaded leaves show the total number of instances ($N_l$) they carry. One thing that we notice even in this pruned version of the resulting decision tree is that all four variables we consider are taking part in the partition process. For example, if we focus on the bottom leftmost leaf in the figure, with $N_l = 1644$, we see that this node contains instances where $\log(ra) \leq 4.91$ `and` $7.5 < dr \leq 21.5$ `and` $0.63 < ea \leq 1.04$ `and` $\log(f) \leq 6.15$.

Once we have the tree structure, our goal is to estimate the Average Treatment Effect (ATE) as presented in Algorithm 1. The treatment effect in our case in a given leaf $l$, $R_l(\boldsymbol{z})$ is defined by the following:

$$R_l(\boldsymbol{z}) = \frac{\sum_{i \in \mathcal{I}_{\boldsymbol{z}}} \mathbb{1}_{S,i}}{|\mathcal{I}_{\boldsymbol{z}}|}\ , \tag{2}$$

where

$$\mathbb{1}_{S,i} = \begin{cases} 1, & \text{if } i \text{ enters the charts} \\ 0, & \text{otherwise.} \end{cases}$$

and $\mathcal{I}_{\boldsymbol{z}}$ is the set of instances within leaf $l$ that have been treated under treatment $\boldsymbol{z}$. The baseline treatment in our study is $\boldsymbol{z}_0$ (see Table 5).

A point worth of note in our approach is that we define the ATE in a different way than the usual form found in the literature (Wang et al. 2015, Austin 2011). In particular, ATE is often defined as the average difference between the treated and the control group. In our case, the difference between success percentages is indicative, but not as descriptive as the rate between the treated and non-treated success percentages. Hence we use the following formula:

$$ATE_{\boldsymbol{z}} = \sum_{l \in \mathcal{L}} \frac{N_l}{\sum_{l \in \mathcal{L}} N_l} \big(R_l(\boldsymbol{z}) / R_l(\boldsymbol{z}_0)\big) \tag{3}$$

In Figure 4a we present the results of our analysis, where for visualization purposes we reduced the three

dimensions to two by concatenating $tl$ and $el$. On the $x$-axis we show the trend ($tl$) and the entropy of the music label ($el$), and on the $y$-axis the total number of releases ($rl$). The shades capture the ATE, as defined by Equation (3). This plot suggests that the stronger effect (around 6 times more likely enter the charts than the baseline) is found at $\boldsymbol{z}_4 = [1, 0, 0]$, suggesting that $\boldsymbol{z}_4$ is the most impactful interaction of the three treatment variables we consider. It is important here to note that isolating $tl$ and arguing that it has the strongest impact on a successful outcome would have been misleading. Other than $Z_4$ we observe that $Z_6 = [1, 1, 0]$ has also a very strong effect (5 times more likely), while $Z_7 = [1, 1, 1]'$ comes third with an ATE of around 4.

These results are very promising since our definitions of "major" labels have been really generous. In particular, so far we have treated all labels than fall into the top 50% of any of the three treatments as "major". However, the really influential labels in our marketplace are these that rank much higher than 50% in these characteristics. To dig a bit deeper, we consider now as top labels those that are ranked in the top 25% of only two of our characteristics: $tl$ and $el$[11]. We ignore all the instances that fall between the bottom 25% and the top 25%. The results are shown in Figure 4b. The effects now become much clearer: if the label's entropy and trend are in the top 25%, then controlling for a series of artists characteristics, a song is 9 times more likely to enter the charts.

Our results show a clear effect between the promoting agencies and the propensity of a song to enter the charts. Given that the marketplace stands to benefit from both diversifying and supporting independent artists, we discuss next a series of suggestions that can steer the platform towards the right direction.

## Discussion

In this section we discuss the implications of our study and we further state its limitations. We further discuss our future research directions on the topic and finally we conclude.

### Marketplace Implications

In this work we analyzed a panel dataset from the major marketplace for electronic music, Beatport.com. Our findings suggest the existence of a strong rich-get-richer phenomenon, where artists that are represented by the top 25% of the music labels in the platform are 9 times more likely to release songs that will enter the charts. From an artist's perspective, our study suggests that it is really important to collaborate with labels that are in the top 25%. However, many new and independent artists avoid collaborating with music labels and

[11] At this level of selection, (top and bottom 25%) we don't consider all three characteristic because of sparsity issues.

they choose to be free and have complete control over their music productions. As a result, assuming that the marketplace is interested in diversifying its rankings to more/new artists, our study pinpoints to two potential directions:

1. Control for the label: The platform should incorporate in the ranking algorithm the prevalence of the music label backing each artist. For example, the platform can employ a Bayesian ranking approach that includes the trending of each label. Such approach could potentially boost artists that are not currently supported by a dominating label.

2. Separate Rankings: The platform could create separate rankingings for independent and new artists.

By employing these proposed actions we firmly believe that Beatport.com will create new opportunities for artists to rise, which will cascade positive effects not only on the marketplace (e.g., increased revenue) but also on the affected community.

## Limitations

Our study exhibits some limitations, mainly originating from our dataset. In particular, due to sparsity issues, we have considered a quantization of the treatment features based on their median value. In a marketplace with abundance of data, the quantization process can be even more granular. For example, we could expect that the top $95^{th}$ quartile would have much stronger effect compared to the one observed within the top $50^{th}$ quartile we utilized in our study. Nevertheless, the results from our analysis are still indicative of the underlying effect and clearly show the promise of similar analysis.

Even though we have used observational data to extrapolate our results, our findings are more than just pure correlations for two reasons. First, we used a state-of-the-art quasi-experimental method and controlled for a series of confounders that might affect the placement of a song in the charts. Second, running a randomized trial is not consistent with the research ethics since the platform would have to randomly assign music labels to different artists/songs. Hence, the selection bias between music-label and artist will always be present.

## Future Directions

Beyond Beatport.com, we wouldn't be surprised if similar effects were to be found in other no barrier-to-entry marketplaces. Having said that, it is important to note that we are not suggesting that our results generalize beyond this specific marketplace. However our presented framework is widely applicable to other similar

platforms. In fact, we are currently creating a long panel dataset with Amazon book rankings. Our intention is to deploy our framework on a completely different setting and study how publishing houses create invisible barriers to independent writers.

Regarding the current version of our work, we intend to expand our dataset in order to study the per-genre effect of music labels. Our current dataset does not allow for an analysis at a finer granularity. We believe that this is an important extension since the per-genre audience is not distributed uniformly; certain genres are mainstream, while others are more underground. Hence, one might expect that the effect of promoting agencies will vary across different genres.

Furthermore, in the current study we have specifically focused on a binary problem, that is, whether the song appeared at the top-100 ranks. Of equal interest is to focus on studying a multi-nary outcome (i.e., top 10, top 20 and so on). Intuition suggests that the top-10 tracks are much more popular than the rest of the top-100. Hence, promoting agencies might (or might not) be strongly associated with a specific part of the ranks as compared to the rest of the top-100.

## Conclusion

To conclude, in this study we focused on understanding how the interplay of different music label characteristics affects the placement of a song in the charts. By collecting and analyzing a unique dataset from a major marketplace of electronic music and by taking into account a set of song features we first deployed an Accelerated Failure Time survival model to estimate the probability of a song to appear in the charts. By doing so, we created a labeled dataset with songs that stochastically have very low propensity of entering the charts, as well as with songs that we already know that have been successful. We then used this dataset and built a multidimensional tree based causal inference framework to study how the interaction of different music label characteristics affect the placement of the song in the charts. We found that certain combinations of the music label's characteristic create a 9-fold increase of a song's probability to enter the charts. Finally, our analysis provides marketplace-specific insights on both the importance of promoting agents on Beatport, but also on potential actionable ways that could help independent artists overcoming these invisible barriers and succeed.

# Figures and Tables

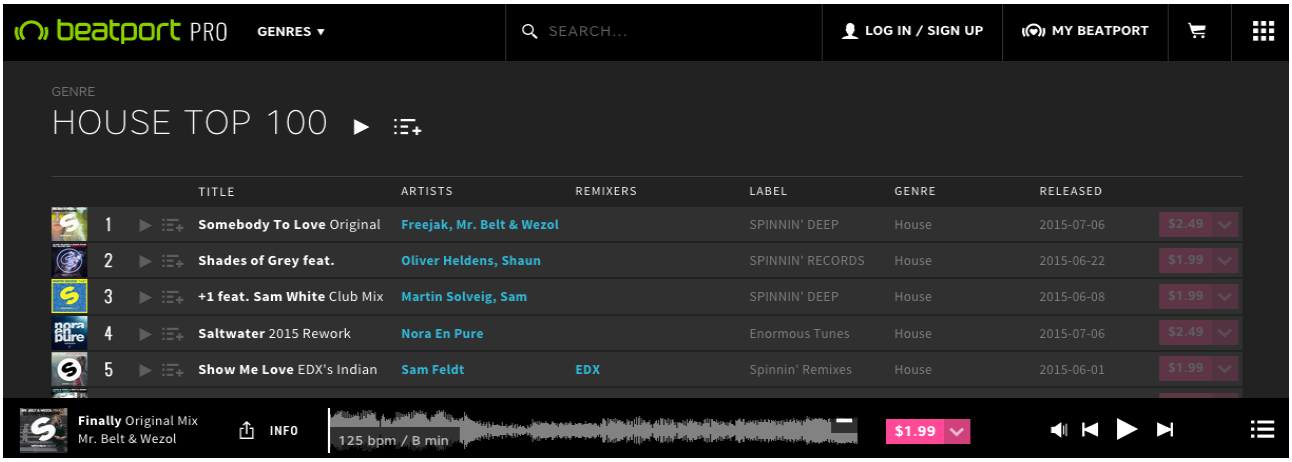



Figure 1: A screenshot from Beatport.com. We see part of the top-100 songs for genre 'House'.

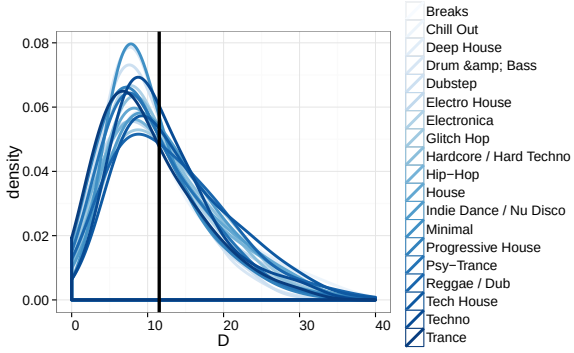



Figure 2: On average it takes $D \approx 12$ days for a song to appear on Beatport's top-100 ranking.

| Variable | Symbol | min | avg | max |
|---|---|---|---|---|
| Label trend | $tl$ | 0 | 0.66 | 39 |
| Label artists | $rl$ | 1 | 1,007 | 16,231 |
| Label entropy | $el$ | 0.21 | 1.79 | 3.16 |
| Artist releases | $ra$ | 1 | 157 | 2,133 |
| Artist entropy | $ea$ | 0.25 | 1.99 | 3.16 |
| Artist followers | $f$ | 1 | 129,478 | 41,779,832 |

Table 1: Descriptive statistics of the variables in our dataset.

| | Artist | Followers |
|---|---|---|
| Top tier | Britney Spears | 41,779,832 |
| | Bruno Mars | 21,700,000 |
| | Pitbull | 20,127,881 |
| | Nicki Minaj | 19,400,000 |
| | will.i.am | 13,167,887 |
| Middle tier | Myon & Shane 54 | 99,371 |
| | Hudson Mohawke | 98,074 |
| | Maceo Plex | 97,682 |
| | Green Velvet | 97,470 |
| | Hot Since 82 | 97,322 |
| Bottom tier | Lanni Johnson | 99 |
| | Agraba | 99 |
| | Phillipo Blake | 99 |
| | What So Not | 99 |
| | Gil Sanders | 98 |

Table 2: Number of Twitter followers of various artists.

| | Label | Releases |
|---|---|---|
| Top tier | Recovery House | 16231 |
| | Armada Music Bundles | 9110 |
| | LW Recordings | 6207 |
| | Club Session | 5809 |
| | Ultra | 4689 |
| Middle tier | Euphonic | 297 |
| | Sportage Digital | 296 |
| | Late Night Records | 295 |
| | Decadencia | 293 |
| | BugEyed Records | 292 |
| Bottom tier | Crash! | 8 |
| | Soak Music | 7 |
| | Cerecs | 6 |
| | Expo Records | 5 |
| | Da South! | 3 |

Table 3: Music Labels' prevalence in the marketplace in terms of the number of released songs $rl$.

| Variable | Coefficient |
|---|---|
| $tl$ | -0.025*** |
| $\log(rl)$ | 0.094 *** |
| $el$ | 0.1842 *** |
| $\log(ra)$ | -0.011 |
| $ea$ | 0.033 |
| $\log(f)$ | -0.048 *** |
| Chi Square | 142.35 on 6 degrees (p-value = 0.000) |

Table 4: Log-normal AFT coefficients. An increase in the entropy of a music label ($el$) or the number of artists under a music label ($rl$), leads to a delay in the entrance of a song on the charts. On the contrary, higher label trending ($tl$) and more popular label – as captured by its number of Twitter followers ($f$)– expedites the appearance of a song on the charts. (Significance codes: '***' 0.001)

**Algorithm 1** Tree-based Causal Inference Algorithm for our scenario.

**Input:** $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$, $Y_i \in \{\text{In Charts, Not in Charts}\}$

1: Learn a tree-based model for estimating $\Pr(\boldsymbol{Z}_i|\boldsymbol{X}_i)$
2: **for** l $\in$ Tree Leaves $\mathcal{L}$ **do**
3: Calculate number of instances, $N_l$
4: Estimate the treatment effect $R_l(t)$, $\forall \boldsymbol{z} \in \mathcal{Z}$
5: **end for**
6: **return** $ATE_{\boldsymbol{z}} = \sum_{l \in \mathcal{L}} \frac{N_l}{\sum_{l \in \mathcal{L}} N_l} \big(R_l(\boldsymbol{z})/R_l(\boldsymbol{z}_0)\big)$

| | $\boldsymbol{z} \in \mathcal{Z}$ | | | Outcome | |
|---|---|---|---|---|---|
| **Symbol** | $tl$ | $el$ | $\log(rl)$ | **Positive** | **Negative** |
| $\boldsymbol{z}_0$ | Bottom 50% | Bottom 50% | Bottom 50% | 141 | 1,434 |
| $\boldsymbol{z}_1$ | Bottom 50% | Bottom 50% | Top 50% | 208 | 5,145 |
| $\boldsymbol{z}_2$ | Bottom 50% | Top 50% | Bottom 50% | 487 | 5,650 |
| $\boldsymbol{z}_3$ | Bottom 50% | Top 50% | Top 50% | 45 | 1,411 |
| $\boldsymbol{z}_4$ | Top 50% | Bottom 50% | Bottom 50% | 195 | 146 |
| $\boldsymbol{z}_5$ | Top 50% | Bottom 50% | Top 50% | 472 | 985 |
| $\boldsymbol{z}_6$ | Top 50% | Top 50% | Bottom 50% | 342 | 334 |
| $\boldsymbol{z}_7$ | Top 50% | Top 50% | Top 50% | 153 | 272 |
| | | | Total | 2,043 | 15,377 |

Table 5: Our treatment set $\mathcal{Z}$ along with the number of positive – songs in the charts – and the number of negative – songs not in the charts – instances for each treatment.

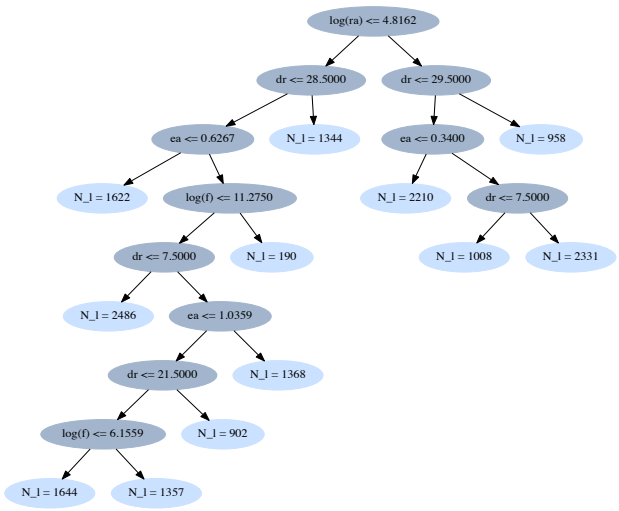
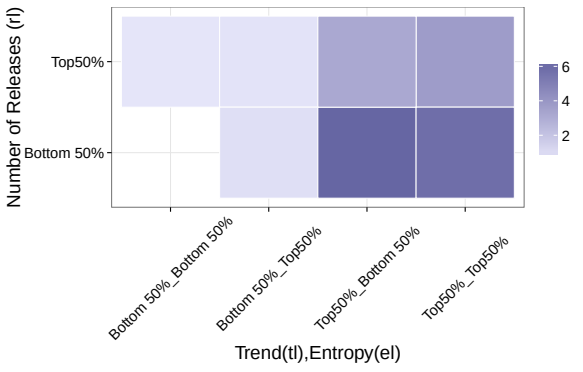
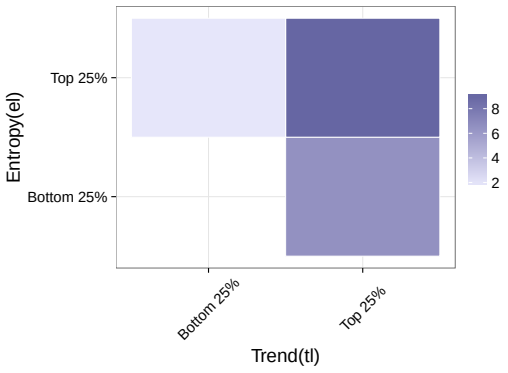



Figure 3: A pruned version of the induced decision tree from our data.




(a) Considering all three music label characteristics, $tl, el$ and $rl$. 5




(b) Considering only $tl$ and $el$, and only the first and fourth quantiles.

Figure 4: The effect of multiple treatments on the song's likelihood of entering the charts.

# Bibliography


Animesh, Animesh, Siva Viswanathan, Ritu Agarwal. 2011. Competing creatively in sponsored search markets: The effect of rank, differentiation strategy, and competition on performance. *Information Systems Research* **22** 153–169. 3

Assmus, Gert, John U Farley, Donald R Lehmann. 1984. How advertising affects sales: Meta-analysis of econometric results. *Journal of Marketing Research* 65–74. 1, 2

Austin, Peter C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* **46** 399–424. 12

Bakos, Yannis, Chrysanthos Dellarocas. 2011. Cooperation without enforcement? a comparative analysis of litigation and online reputation as quality assurance mechanisms. *Management Science* **57** 1944–1962. 4

Bhattacharjee, Sudip, Ram D Gopal, Kaveepan Lertwachara, James R Marsden, Rahul Telang. 2007. The effect of digital sharing technologies on music markets: A survival analysis of albums on ranking charts. *Management Science* **53** 1359–1374. 4

Blattberg, Robert C, Abel P Jeuland. 1981. A micromodeling approach to investigate the advertising-sales relationship. *Management Science* **27** 988–1005. 1, 2

Bolton, Gary E, Elena Katok, Axel Ockenfels. 2004. How effective are electronic reputation mechanisms? an experimental investigation. *Management science* **50** 1587–1602. 4

Bradlow, Eric T, Peter S Fader. 2001. A bayesian lifetime model for the top 100 billboard songs. *Journal of the American Statistical Association* **96** 368–381. 4

Chen, Hailiang, Prabuddha De, Yu Jeffrey Hu. 2015. It-enabled broadcasting in social media: An empirical study of artists activities and music sales. *Information Systems Research* **26** 513–531. 3

Chevalier, Judith A, Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research* **43** 345–354. 4

Clarke, Darral G. 1976. Econometric measurement of the duration of advertising effect on sales. *Journal of Marketing Research* 345–357. 1, 2

Cleves, Mario. 2008. *An introduction to survival analysis using Stata*. Stata Press. 8

Cox, David Roxbee, David Oakes. 1984. *Analysis of survival data*, vol. 21. CRC Press. 8

Danaher, Brett, Yan Huang, Michael D Smith, Rahul Telang. 2014. An empirical analysis of digital music bundling strategies. *Management Science* **60** 1413–1433. 4

Dellarocas, Chrysanthos. 2006. Reputation mechanisms. *Handbook on Economics and Information Systems* 629–660. 4

Dewan, Sanjeev, Jui Ramaprasad. 2012. Research note-music blogging, online sampling, and the long tail. *Information Systems Research* **23** 1056–1067. 3

Dewan, Sanjeev, Jui Ramaprasad. 2014. Social media, traditional media, and music sales. *Mis Quarterly* **38** 101–121. 3

Dhar, Vasant, Elaine A Chang. 2009. Does chatter matter? the impact of user-generated content on music sales. *Journal of Interactive Marketing* **23** 300–307. 3

Fleder, Daniel, Kartik Hosanagar. 2009. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management science* **55** 697–712. 3

Ghose, Anindya, Panagiotis G Ipeirotis, Beibei Li. 2014. Examining the impact of ranking on consumer behavior and search engine revenue. *Management Science* **60** 1632–1654. 3

Ghose, Anindya, Sha Yang. 2009. An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science* **55** 1605–1622. 3

Grambsch, Patricia M, Terry M Therneau. 1994. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81** 515–526. 8

Kokkodis, Marios, Panagiotis G Ipeirotis. 2015. Reputation transferability in online labor markets. *Management Science* . 4

Koster, Alexis. 2011. The emerging music business model: back to the future? *Journal of Business Case Studies (JBCS)* **4** 17–22. 1

Lee, Jonathan, Peter Boatwright, Wagner A Kamakura. 2003. A bayesian model for prelaunch sales forecasting of recorded music. *Management Science* **49** 179–196. 4

Lu, Xianghua, Sulin Ba, Lihua Huang, Yue Feng. 2013. Promotional marketing or word-of-mouth? evidence from online restaurant reviews. *Information Systems Research* **24** 596–612. 3

Nelson, Phillip. 1970. Information and consumer behavior. *The Journal of Political Economy* 311–329. 3

Resnick, Paul, Richard Zeckhauser, John Swanson, Kate Lockwood. 2006. The value of reputation on ebay: A controlled experiment. *Experimental economics* **9** 79–101. 4

Salganik, Matthew J, Peter Sheridan Dodds, Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science* **311** 854–856. 3

Shadish, William R.., Thomas D Cook, Donald Thomas Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning. 10

Strobl, Eric A, Clive Tucker. 2000. The dynamics of chart success in the uk pre-recorded popular music industry. *Journal of Cultural Economics* **24** 113–134. 4

Wang, Pengyuan, Wei Sun, Dawei Yin, Jian Yang, Yi Chang. 2015. Robust tree-based causal inference for complex ad effectiveness analysis. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 67–76. 11, 12