# Split-door criterion for causal identification: Natural experiments with testable assumptions

**Amit Sharma**[1,*], **Jake Hofman**[1], **and Duncan Watts**[1]

[1]Microsoft Research, New York
[*]amshar@microsoft.com

## ABSTRACT

Unobserved or unknown confounders complicate even the simplest attempts to estimate the effect of one variable on another using observational data. While there are a number of different approaches to eliminate confounds in the causal inference literature, each has its own set of assumptions, many of which are difficult to verify precisely because they involve statements about variables that, by definition, cannot be measured. In this paper we investigate a particular scenario that both permits causal identification in the presence of unobserved confounders and has explicitly testable assumptions stated only in terms of observable variables. Specifically, we examine what we call the *Split-door* setting, when the effect variable can be split up into two parts: one part that is potentially affected by the cause, and another other that is independent of it. We show that when both of these variables are caused by the same (unobserved) confounders, the problem of identification reduces to that of testing for independence among observed variables. We discuss various situations in which Split-door variables are commonly recorded in both online and offline settings, and demonstrate the method by estimating the causal impact of Amazon's recommender system, obtaining similar—but more precise—estimates than past studies.

## 1 INTRODUCTION

In domains ranging from the social sciences, biomedical sciences to business, there are a number of applications where data is available for two quantities of interest and our goal is to estimate the causal effect of one over other. For example, one may be interested in the effect of education on future earnings[1], the impact of genetic markers on prevalence of certain diseases[2], or the impact of ads or recommendations on purchases in an e-commerce store[3].

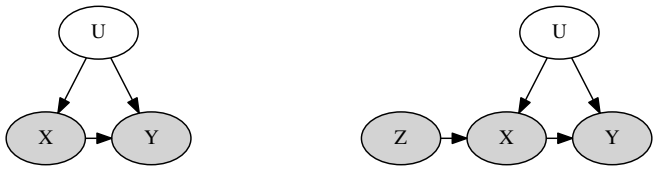
Unlike problems where even the direction of causality is unknown[4], in these problems, domain knowledge allows us to clearly identify which one of the variables is the cause and which one the effect, or outcome. Figure 1 shows the canonical class of causal inference problem that we study in this paper, where X is the cause and Y is its effect. $U$ refers to all the unobserved and potentially unknown *confounders* that could be the common cause of X and Y. For example, both selection into a job training program and future earnings may be driven by intrinsic skill of an individual, both genes and disease may be impacted by environmental factors, and customers' purchases and the ads they are exposed to may depend on their interest in certain products. The presence of such unobserved confounders makes the task of isolating the causal effect of X on Y a difficult one.

A common approach is to assume that the effect of unobserved confounders is negligible compared to the observed variables in U. Under such a *selection on observables* assumption[5], we can condition on observed confounders to estimate the effect of X on Y when confounders stay constant. In the language of graphical models, this strategy is based on the backdoor criterion[6] and can be implemented by a variety of methods, including stratification and matching. However, in most practical problems, it is hard to justify that all of the important confounders will be observable. For instance, in our examples above, it is almost impossible to accurately measure intrinsic skill, environmental factors or customer interest in certain products.

The ideal solution for dealing with unobserved confounders is to run a randomized experiment where X is manipulated independently of U, but such experiments are not always possible in practice. Given these limitations, researchers in the social sciences have developed alternative approaches for identifying causal effects where one looks for naturally occurring variation in the data that is arguably random. The hope is that such variation, known as a *natural experiment*, can serve as a proxy for an actual randomized experiment.

Suppose, for instance, that our goal is to estimate the impact of recommendations on an e-commerce store. The ideal experiment would be to turn the recommendation system on or off uniformly at random for different users. In expectation, such a randomization scheme nullifies the effect of confounders. Without the luxury of conducting an experiment, the crux of a

This is a preprint draft. This version is made available for presentation and receiving early feedback at the 2016 UAI Workshop on "Causation: Foundation to Application".




**(a)** U confounds the effect of X on Y

**(b)** Estimation with Z as an instrument

**Figure 1.** Graphical model for the canonical problem in causal inference. We wish to estimate the effect of X on Y. U represents all the unobserved (and unknown) confounders that commonly cause both X and Y.

natural experiment lies in identifying an external event that brings a sizeable change in the data distribution. For example, one might look for a book that was featured on Oprah's book club, bringing a large and sudden change in the number of customers who visit the book's webpage on Amazon[7]. Assuming that the additional people who visit during this event are effectively a random sample of customers, one can estimate the causal effect of the recommender by measuring the change in sales to the products recommended alongside the featured book, arguing that these sales would not have happened in the absence of the recommender.

However, the core assumptions underlying the validity of many natural experiments are difficult to verify. Figure1(b) shows an extension of the canonical graphical model for a natural experiment, where an additional observed variable Z controls the variation in X. Here $Z$ can be considered as an *instrumental variable* that helps in identification of causal effect[8]. Continuing from our example above, causal identification relies on two assumptions: that Oprah's followers are a nearly random sample from the population of customers ($Z \perp\!\!\!\perp U$) and that they are not more interested in recommendations on the page than any other customer (Z does not cause Y directly). Typically, researchers use the domain knowledge to justify these assumptions, but such qualitative justifications are far from satisfactory, especially since validity of the causal estimate hinges on them.

Further, coming up with such arguably random variations requires creative ingenuity on the part of a researcher. This severely limits the generalizability of the results from a natural experiment, and applicability of the method to other domains. In practice, most natural experiments rely on just a handful of sources for random variation, such as lotteries, the weather or a sudden, large event[9].

In this paper, therefore, we present a data-driven causal identification strategy that can be used to search for and validate natural experiments from observational data. Our strategy requires that the outcome, $Y$ be split into two constituents: one that is affected by $X$ and the other that is independent of $X$. This leads to the causal graphical model as shown in Figure 2, where $Y_D$ refers to the "direct" constituent of $Y$ that is not affected by $X$. Whenever such fine-grained data on $Y$ is available, we show that it is possible to identify the causal effect, and propose a simple, scalable algorithm for doing so. Because this strategy depends on availability of a split set of variables for $Y$, we call it the *Split-door* criterion for causal identification.

Intuitively, the Split-door criterion works by generalizing the notion of large and sudden variations in $X$ that natural experiments typically exploit. By automatically searching for experiments, instead of coming up with them, the Split-door strategy is able to analyze a larger fraction of the overall data, thus increasing the chances of obtaining generalizable conclusions. Critically, it also tests that each of the variations in $X$ is not caused by the same confounders that bias the effect between $X$ and $Y$.We show that whenever $Y_D$ is statistically independent of $X$, it is possible to identify the causal effect of $X$ on $Y$.

The requirement for fine-grained data on $Y$ may seem as a stringent one, but in many domains of interest, $Y$ is commonly recorded in terms of its different constituents. At least in the digital world, such constituent data is usually recorded for variables and easy to access from system logs, in applications such as recommendation, advertising and web traffic. For example, for ads or recommendations in online systems, click-throughs from recommendations and direct webpage visits are separately recorded and can be accessed from the system logs. Further, as the size of available datasets and the number of variables in them increase at a fast rate, such a data-driven approach promises to be of value.

Finally, we demonstrate an application of the Split-door criterion by using it to estimate the causal impact of recommender systems. We find that observational metrics such as the click-through rate (CTR) overestimate the actual effect by over 200%. Our causal estimates are similar to past work using a natural experiments approach[3].

## 2 RELATED WORK

The present work builds upon past research in causal identification criteria, validation of natural experiments and building data-driven algorithms for causal estimation. We discuss each of these in turn.

### 2.1 CAUSAL IDENTIFICATION CRITERIA

There is a rich body of work on using graphical models to identify when causal effect can be estimated using observational data alone. Although do-calculus allows us to test identification for any general graph, two of the major identification criteria are backdoor and frontdoor criteria[6]. The backdoor criterion conditions on known common causes between X and Y to estimate the causal effect. The frontdoor criterion depends on knowledge of the mechanism by which effect is transferred from X to Y. Knowing the mediation variable between X and Y allows us to identify the causal effect.

However, in the absence of observed confounders or mediating variables, there are no obvious strategies that guarantee identification of the causal effect. We isolate a specific, common case of a graphical model and propose the Split-door criterion to identify the causal effect.

### 2.2 VALIDATION OF NATURAL EXPERIMENTS

Our work is closely related to natural experiments in the social sciences[8]. Natural experiments rely on finding subsets of observational data in which causal effect can be identified. The key requirement is that criterion used for subsetting the data should not affect the outcome directly. This is known as the *exclusion* restriction. Satisfying the exclusion restriction guarantees valid identification of the causal effect for the subset data, but we need another condition, *as-if-random*, to make sure that the estimate on the subset data also generalizes to the entire dataset. In other words, the as-if-random restriction requires that the data subset selection itself should be independent of the common causes of X and Y. In sum, these restrictions stipulate that the outcome of interest should be independent of the data subsetting procedure.

Critically, however, both of these assumptions depend on properties of the unknown confounders and are largely believed to be untestable[10]. Therefore, most of the empirical research follows a design-based approach[11], where theoretical justifications drive the validity of causal identification.

As we will show, having access to an additional variable can directly let us test for the subsets of data where the treatment X is not confounded with Y. Each of these subsets can be thought of a hypothetical experiment where the change in X is independent of the confounding causes for X and Y.

### 2.3 DATA-DRIVEN NATURAL EXPERIMENTS

The Split-door criterion can be thought of as a generalization of recent work on searching for instrumental variables in observational data. Taking advantage of the fine-grained data available through online systems, past work searches for data subsets where X experiences a large and sudden spike in its value, while one of the split-up Y remains constant[3]. These subsets are meant to capture events resembling "exogenous shocks", that act as instruments in driving the value of $X$ up.

In this work, we generalize the notion of shocks by weakening the conditions required for valid identification. Intuitively, a shock to $X$ and constancy in split-up direct $Y$ conveys that $X$ and direct Y are not affected by each other. This criterion turns out to be a special case of independence between $X$ and direct $Y$. We show next how independence between $X$ and direct $Y$ allows for identification of causal effect.

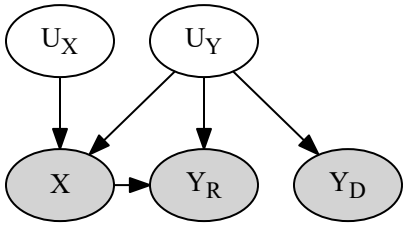



**Figure 2.** Graphical model for the Split-door criterion. The goal is to estimate the causal effect of $X$ on $Y$.

## 3 SPLIT-DOOR IDENTIFICATION CRITERION

We first describe the core assumptions behind the proposed Split-door identification criterion. Under these assumptions, we demonstrate that the criterion can be used to estimate causal effect of $X$ on $Y$, first using graphical models (Section 3.1) and then using parametric regression (Section 3.2). Based on this reasoning, we propose an algorithm for identifying causal effect using the Split-door criterion in Section 3.3.

### 3.1 SPLIT-DOOR THROUGH A GRAPHICAL MODEL

Figure 2 shows the causal graphical model that we assume generates the observed data. Here $X$ represents the cause and $Y$ its effect. $Y$ can be broken down into its constituents: $Y_R$ represents the part that could be affected by $X$, and $Y_D$ represents the part that is not affected by $X$ at all. We assume $Y = Y_R + Y_D$, where $Y_R$ is the *referred* part and $Y_D$ is the *direct* part. $U_Y$ represents the unobserved causes for $Y$, some of which may also be common causes for $X$, thus the arrow from $U_Y \rightarrow X$. $U_X$ represents the unobserved causes for $X$ that are *independent* of $U_Y$. Both $U_X$ and $U_Y$ can be a combination of many variables, some observed and some unobserved. These unobserved and potentially unknown variables $U_X$ and $U_Y$ make it difficult to estimate the effect of $X$ directly from observational data.

As an example, suppose that a company routinely sends discount coupons for one of its products to some of its customers. The company may want to know the return on giving away discounts. In this example, $X$ would be the number of customers that are sent a discounted offer, $Y_R$ would be the customers among them who used the discount to purchase the product, and $Y_D$ would be the number of customers who bought the product through other channels, without a discount. Our goal is to find the causal effect of giving away discounts on product purchases: what would have happened if company had cut its discounts to half the people, or did not give away discounts at all?

That is, we are interested in the effect of X on Y, when they are generated according to a causal model as in Figure 2. The problem becomes tricky because unobserved demand $U_Y$ for the product may confound the effect of $X$ on $Y_R$. To identify the causal effect, we use data for purchases of the product ($Y_D$) that were not referred through a discount coupon, but are also expected to be jointly caused by the (unobserved) demand $U_Y$ for the product.

An equivalent way of writing the causal model is through specifying the three structural equations:

$$x = g(u_x, u_y, \varepsilon_x) \tag{1}$$
$$y_r = f(x, u_y, \varepsilon_{yr}) \tag{2}$$
$$y_d = h(u_y, \varepsilon_{yd}) \tag{3}$$

where $\varepsilon_x$, $\varepsilon_{yr}$, and $\varepsilon_{yd}$ are mutually independent variables that correspond to modeling error and statistical variability. Because $Y_D$ and $Y_R$ are two additive constituents of the same variable $Y$, we can assume that $U_Y$ affects both $Y_D$ and $Y_R$. In other words, we assume that $Y_D$ is not independent of $U_Y$ and the function $h$ depends on $U_Y$.

More precisely, we make the following two assumptions to identify the causal effect.

**Assumption 1** [*Connectedness*]: *Any unobserved confounder $U_Y$ that causes both $X$ and $Y_R$ also causes $Y_D$ and the causal effect of such $U_Y$ on $Y_D$ cannot be zero.*

This assumption also applies to all sub-components of $U_Y$. That is, we assume that any unobserved sub-component $U_{Yi}$ of $U_Y$ that causes both $X$ and $Y_R$ also causes $Y_D$ and the causal effect of such $U_{Yi}$ on $Y_D$ cannot be zero.

**Assumption 2** [*Faithfulness*]: *If $X$ and $Y_D$ are statistically independent, then they are also causally independent in the graphical model of Figure 2.*

That is, the observed data distribution is *faithful*[12] to the causal graphical model in Figure 2. The above assumption serves to rule out incidental equality of parameters in the structural equations that may lead to the observation that $X$ and $Y_D$ are statistically independent.

In our example with discount coupons above, *Assumption 1* implies that any sub-component of the product demand $U_Y$ that causes both $X$ and $Y_R$ also affects the direct purchases $Y_D$. This is reasonable to assume, since $Y_D$ and $Y_R$ are purchases of the same product. *Assumption 2* rules out the possibility of an (unlikely) event where the effect of the unobserved demand $U_Y$ (or its sub-components) cancels out exactly in a way such that $X$ and $Y_D$ become statistically independent.

Using the above two assumptions, and properties of the graphical model, we can show that statistical independence of $X$ and $Y_D$ ensures that $X$ is not confounded by $U_Y$.

**Theorem 1.** *If $X$, $Y_R$ and $Y_D$ are three random variables generated by the process shown in Figure 2, and the Connectedness and Faithfulness assumptions hold, then $X \perp\!\!\!\perp Y_D$ implies that the effect of $X$ on $Y$ is not confounded by $U_Y$.*

*Proof.* The proof can be completed directly from Figure 2. $X \perp\!\!\!\perp Y_D$ implies that the causal effect of $U_Y$ on $Y_D$ and $X$ somehow cancels out on the path $X \leftarrow U_Y \rightarrow Y_D$. By *Assumption 2*, the joint data distribution is faithful to the causal graph, so this can only happen if $U_Y$ is constant (and thus *blocks* the path), or one of the edges exists trivially (does not have a causal effect).

(i) By the backdoor criterion[6], if $U_Y$ is constant, then $X$ and $Y$ are unconfounded, because the only backdoor path between $X$ and $Y$ contains $U_Y$ on it.

(ii) Using *Assumption 1*, $U_Y$ has a non-zero effect on $Y_D$. Then, the only alternative is that $X \leftarrow U_Y$ edge does not exist.

In both (i) and (ii) scenarios, there are no backdoor (confounding) paths from $X$ to $Y$ and thus the effect of $X$ on $Y$ is unconfounded. □

Without the confounding due to $U_Y$, the observational estimate is also the causal estimate: $P(Y|do(X=x)) = P(Y|X=x)$.

### 3.2 SPLIT-DOOR THROUGH PARAMETRIC REGRESSION

Another way to interpret the Split-door criterion is by using linear regression equations. Although the causal effects among variables may not be necessarily linear, presenting the analysis in a parametric form clarifies the intuition behind the criterion.

Let $x$, $y_r$ and $y_d$ be the observed values for $X$, $Y_R$ and $Y_D$ respectively. Then, using the structural equations above, we can write:

$$x = \eta u_x + \gamma_1 u_y + \varepsilon_x \tag{4}$$

$$y_r = \beta x + \gamma_2 u_y + \varepsilon_{yr} \tag{5}$$

$$y_d = \gamma_3 u_y + \varepsilon_{yd} \tag{6}$$

where $\varepsilon_x$, $\varepsilon_{yr}$ $\varepsilon_{yd}$ are independent errors in the regression equations. Independence of $X$ and $Y_D$ implies that:

$$\begin{aligned} Cov(X,Y_D) &= E[XY_D] - E[X]E[Y_D] = 0 \\ &= E[(\eta u_x + \gamma_1 u_y + \varepsilon_x)(\gamma_3 u_y + \varepsilon_{yd})] \\ &- E[\eta u_x + \gamma_1 u_y + \varepsilon_x]E[\gamma_3 u_y + \varepsilon_{yd}] \\ &= \gamma_1\gamma_3 E[U_Y.U_Y] - \gamma_1\gamma_3 E[U_Y]E[U_Y] \end{aligned}$$

Assuming that $y_d$ is affected by $U_Y$ (and therefore $\gamma_3$ is not 0), the above can be zero only if $\gamma_1 = 0$, or if U is constant (and thus $E[U_Y]E[U_Y] = E[U_Y.U_Y]$). In both cases, $X$ becomes independent of $U_Y$ and the following regression can be used as an unbiased estimator for the effect of $X$ on $Y_R$.

$$y_r = \beta x + \varepsilon'_{yr} \tag{7}$$

### 3.3 ALGORITHM FOR FINDING NATURAL EXPERIMENTS

The above two results point to a simple algorithm for finding natural experiments in observational time-series data. We split the data into equally-spaced time-intervals $\tau$, in a way that each interval has enough data points to reliably estimate the joint probability distribution $P(X, Y_R, Y_D)$.

Then, for each time interval $\tau$,

- Determine whether $X$ and $Y_D$ are independent using an empirical independence test.
- If $X \perp\!\!\!\perp Y_D$, then use the observed conditional probability $P(Y|X=x)$ to estimate the causal effect in the interval $\tau$.
- Average over all time-intervals where $X \perp\!\!\!\perp Y_D$ to obtain the mean causal effect of X on Y.

Assuming we have enough data, we can test for independence by comparing the empirical mutual information to zero. Assuming the null hypothesis that $X \perp\!\!\!\perp Y_D$ (and hence mutual information is zero), we can determine whether the empirical deviation from zero is significant using a permutation test with a pre-chosen $\alpha = 0.05$ significance level.

Note that the estimate obtained is not the Average Causal Effect (ACE), because we are selecting the time-intervals in a non-randomized fashion. This implies that while any estimated causal effect is a valid estimate for the time-intervals where Split-door criterion is satisfied, it may not generalize to the full dataset. This is not a problem with Split-door criterion *per se*, but a general limitation of natural experiments. Nevertheless, compared to past methods for natural experiments that depend on single-source events[8,9,11], the criterion allows us to compute causal estimate over a larger sample of data, thereby increasing the generalizability of the estimate.


**Figure 3.** Screenshot of recommendations shown for each focal product on Amazon.com

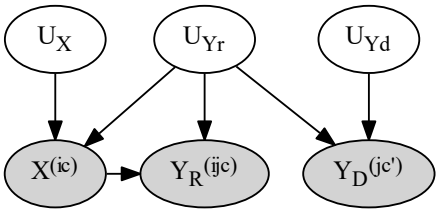



**Figure 4.** Graphical model for estimating the impact of recommendations on Amazon.com.

## 4 APPLICATION: IMPACT OF A RECOMMENDER SYSTEM

We now show a practical application of the Split-door criterion, by using it to estimate the causal impact of a recommender system.

Recommender systems[13] are ubiquitous, providing suggestions for what to buy, watch, read or do next. For example, online retailer Amazon.com provides recommendations for other products on each product's webpage, as shown in Figure 3. In the example screenshot, the book "California" is the *focal* product and the recommendations shown on its webpage are the *recommended* products. We are interested to know how much traffic the recommender system causes, to evaluate its marginal impact and also to propose metrics for improvements to the recommender system.

Evaluating the impact of recommendations is a tricky problem, because it requires estimating the counterfactual: what would have happened to a website's traffic or sales in case there were no recommendations? A naive way to estimate impact would be to simply count the number of click-throughs from the recommendations shown. However, some of these click-throughs may just be due to *convenience* and people might already know about the recommended item or may have found out about it through other channels such as search. This is even more likely with recommendations, because they are designed to be similar to the item they are shown against, hence increasing the chances that a user already knows about them.

### 4.1 BUILDING THE CAUSAL MODEL

The above discussion suggests that there is some unobserved common demand for both the focal product and the recommended product, which biases observational estimates such as click-through rate (CTR) on recommendations. Figure 4 presents the graphical model for this problem, showing how common demand acts as a confounder between visits to the focal and recommended product.

- $X$ denotes a visit by a customer $c$ to the focal product $i$'s webpage.
- $Y_R$ denotes whether the customer $c$ clicked on the recommendation for product $j$ on product $i$'s webpage.

- $Y_D$ is the number of direct visits to product $j$ by other customers $c'$ in the same time period. These could be visits to $j$ from Amazon's search page, or through a direct visit to $j$'s webpage URL.
- $U_{Y_r}$ represents unobserved demand of customer $c$ for product $j$, some of which may be common with demand for product $i$. This demand for the product $j$ is also connected to the direct visits by other customers $c'$. Therefore, we add edges $U_{Y_r} \rightarrow X$ and $U_{Y_r} \rightarrow Y_D$.
- $U_{Y_d}$ represents the unobserved demand for product $j$ among the customers $c'$ who visit product j directly, that is independent of—not explained by—$U_{Y_r}$.
- $U_X$ represents the unobserved demand for product $i$ that is independent of $U_{Y_r}$.

The sum of $Y_R$ and $Y_D$ is the total number of visits to product $j$. We wish to estimate the causal effect of $X$ on $Y$. As in the graphical model for the Split-door criterion we introduced in Figure 2, $Y$ is split up into two constituents: $Y_R$ and $Y_D$, and $X$ directly causes only one of them. Note that Figure 4 differs from the Split-door model by separating out $U_Y$ into $U_{Y_R}$ and $U_{Y_D}$ . This is because customers who visit product $j$ through recommendation from product $i$ are not the same as those who will visit $j$ directly. In other words, we expect a different population of users to visit product $j$ directly.

However, since both $Y_R$ and $Y_D$ correspond to visits to the same product $j$, we assume that that the unobserved demand $U_{Y_R}$ that confounds $X$'s effect on $Y_R$, also partly causes direct visits $Y_D$ to product $j$. This is represented in the graphical model by adding the edge $U_{Y_R} \rightarrow Y_D$. Further, it is reasonable to assume that the unobserved demand $U_{Y_R}$ affects both $X$ and $Y_D$ in the same direction: the effect is either increasing or decreasing for both. This is because the product $j$ is a recommendation for $i$, and therefore demands for the two products are expected to be positively correlated to each other. Therefore, the Faithfulness assumption is satisfied, because the effect of $U_{Y_R}$ cannot be canceled out on the path $X \leftarrow U_{Y_R} \rightarrow Y_D$ if the effects of $U_{Y_R}$ (and any of its sub-components) on $X$ and $Y_D$ are in the same direction. Given this additional information, the same reasoning from Section 3 allows us to derive $X \perp\!\!\!\perp Y_D$ as a valid identification criterion.

### 4.2 DATA FROM WEB BROWSING LOGS

We obtain anonymized browsing logs of users who have installed the Bing Toolbar and have consented to provide their data. These logs are for nine months, from September 2013 to May 2014, spanning 23.4 million visits by 2.1 million users to 1.38 million unique products. We restrict our analysis to products with at least 10 page views on any single day during the nine month period, leaving us with 22K products.

Amazon.com provides detailed parameters in each URL that is accessed, allowing us to identify the precise source of each visit: which ones are due to recommendation, which ones due to search, and so on. Based on which product people were browsing before a recommendation visit, we are also able to identify the focal product for each recommended product. Thus, we can reconstruct the page visits for each session for a user. More details about the dataset can be obtained in[3].

We consider "Customers who bought this also bought" recommendations, because these recommendations are the most common and are shown on product pages from all categories.

### 4.3 FINDING NATURAL EXPERIMENTS

We now apply the Split-door criterion to search for natural experiments. The split-door method works by searching for cases when observed $X$ and $Y_D$ are statistically independent. With the assumption that demands for $Y_D$ and $Y_R$ are correlated, that implies that $U_{Y_R}$ and $X$ are not directly connected by a causal relationship. Therefore, the causal effect of X on Y is identified using the observational probability $P(Y_R|X)$.

We define time interval $\tau$ as equal chunks of 15 days. The choice of 15 days is guided by empirical considerations: we require a time period over which we will have enough data for estimation, and over which Amazon's recommendations are expected to stay constant.

For each $\tau = 15$ days period:

1. Construct $X$ as the number of visits to a focal product on each day, and $Y_R$ as the number of click-throughs to a recommended product $j$. Record the total direct visits ($Y_D$) to product $j$ on the same day.
2. For each (focal product, recommended product) pair, check whether $X$ is independent of $Y_D$ using a suitable empirical test.

Since we are interested in estimating causal exposure from the recommender system, we only consider a customer's first visit to a product as the source of exposure. Therefore, the same user cannot discover a product $j$ both through click-through on a recommendation and searching directly.

Here we filter out any time interval where $Y_D$ is exactly constant (because that will satisfy empirical independence conditions trivially).

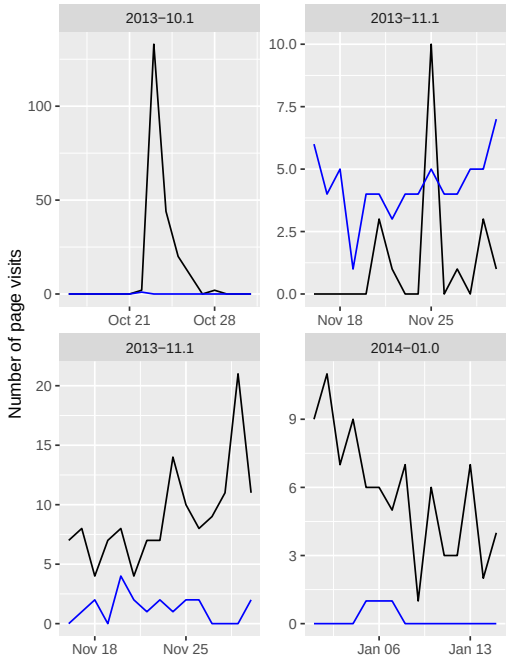



**Figure 5.** Examples of visit timeseries for focal (in black) and recommended (in blue) products that are accepted by the Split-door criterion.

- If $X$ is empirically independent of $Y_D$, compute observed CTR = $(\sum_{t=1}^{\tau} Y_R)/(\sum_{t=1}^{\tau} X)$ during this period as the causal estimate for CTR.

3. For each focal product, summing up the causal CTR estimate over all recommended products provides the aggregated causal CTR.

Averaging this estimate over all time periods allows to compute a mean causal effect for each focal product.

Although estimating mutual information is one of the most direct ways of determining independence, constraints in our empirical data call for a modification of the approach. Each subset contains at most 15 values for testing independence between $X$ and $Y_D$. For such small sample sizes, mutual information estimation can be heavily biased. Therefore, we employ an alternative test, Fisher's exact test, that has better properties in the small sample regime[14].

To do so, we first define a threshold of independence based on the significance level of Fisher's independence test. The independence test works by assuming a null hypothesis that $X$ and $Y_D$ are independent and then proceeds to reject the hypothesis based on a significance level $\alpha$. Here we are interested in finding the cases where $X$ and $Y_D$ *are* independent, so we define the threshold of independence as $1 - \alpha$. As the threshold decreases, our confidence of not rejecting only the cases where $X$ and $Y_D$ are independent increases.

Figure 5 shows some examples of $X$ and $Y_D$ pairs that the Split-door criterion accepts at a threshold of 0.05. The first panel simply shows a timeseries closest to an exogenous shock, but the other panels show the general nature of the split-door criterion: allowing varying but independent variations between $X$ and $Y_D$. Conversely, the timeseries that are rejected by the Split-door criterion are shown in Figure 6. Each of the $X$ and $Y_D$ timeseries show a correlation pattern between them, leading the criterion to reject these pairs for computing causal estimates. Using the Split-door criterion, we obtain a total of 23k natural experiments—(focal product, recommended product) pairs.

### 4.4 ESTIMATING CAUSAL CTR

If we compute a naive estimate over all products, we obtain the mean observational CTR as 9.6%. Our goal is to use the natural experiments identified by the Split-door criterion to obtain a better estimate of the impact of recommender systems.

Figure 7 shows the causal CTR ($\rho$) as a function of the $(1 - \alpha)$ for the independence test between $X$ and $Y_D$. We find that as we decrease the threshold, the value of average causal CTR tends to stabilize. At the lowest threshold (*0.001*), causal CTR is 3.4%. At the other end, at a threshold of 0.2, causal CTR is 3.8%. These estimates show that naive observational CTR is a

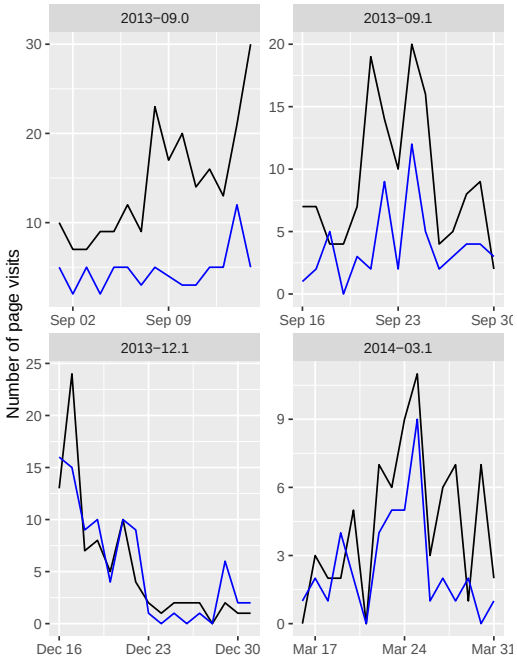



**Figure 6.** Examples of visit timeseries for focal (in black) and recommended (in blue) products that are rejected by the Split-door criterion.

huge overestimate, confirming prior findings using a different approach on the same dataset where average causal CTR was reported to be roughly 3%[3].

We can also compare causal CTR between different product categories on Amazon. Figure 8 shows the variation of $\rho$ with the most popular categories, at an independence threshold of 0.05. We see a big variation in the naive estimate, ranging from 14% on *e-Books* to 6% on *Personal Computer* and *Movies*. However, when we use the Split-door criterion to compute estimates, causal CTR for all product categories lies below 5%. These results indicate that the naive estimate overestimates the causal impact by 200-400% across different product categories.

There are two clear advantages to the Split-door approach. First, we are able to study a bigger fraction of products than with natural experiments that depend on single-source variations[9] or mining shocks in observational data[3] . On the same dataset, the shock-based method by Sharma et al. identified natural experiments on 4k products, while the Split-door approach finds natural experiments for over 12k products. Second, the method yields a more interpretable tuning parameter— the p-value for independence between $X$ and $Y_D$—compared to the parameters used in previous work.

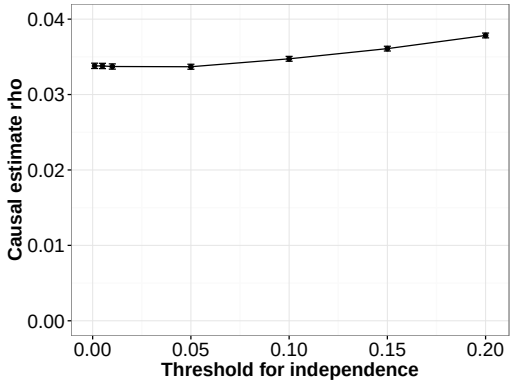



**Figure 7.** Variation of causal CTR ($\rho$) with $1-\alpha$ value of Fisher's independence test, where $\alpha$ is the significance level of the test.

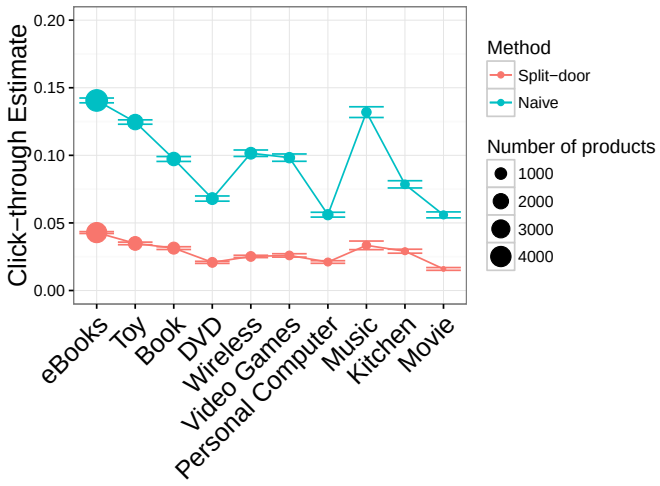



**Figure 8.** Comparison of causal CTR with Naive CTR for products that are satisfy the Split-door criterion. Top 10 product categories based on the number of products in the Split-door sample.

### 4.5 GENERALIZATION ISSUES

While the Split-door criterion yields valid estimates of the causal impact of recommendations for the time periods where $X$ and $Y_D$ are independent(subject to robust empirical detection of independence), care must be taken to extrapolate these estimates to all products on Amazon.com.

Fortunately, the Split-door criterion found valid natural experiments for nearly a third of all focal products in our dataset. Further, the distribution of products in the Valid Split-door subset closely resembles the overall distribution of products with reasonable popularity—with $\geq$10 visits on any one day—as shown in Figure 9. While the as-if-random assumption may not be necessarily valid, these results suggest that the Split-door criterion allows us to study a representative sample of products on Amazon, at least according to number of products and visits by product category.

## 5 DISCUSSION

We have presented a method for computing causal effect of a variable $X$ on another variable $Y$ whenever we know additional variable $Y_D$ which follows some testable conditions. This method can be applied in any domain where such a $Y_D$ is available.

We now present some example domains where the Split-door criterion may be applied. We also discuss the key assumptions and limitations of the criterion that must be checked prior to applying it.

### 5.1 APPLICATIONS

The most direct application is in online systems, such as recommendation or ad systems. In these systems, $Y_D$ has a natural interpretation of "direct traffic", or any traffic that was not caused by a particular recommendation or ad. We applied the Split-door criterion to Amazon's recommendation system, but it can be similarly applied to other recommendation systems, search engine ads or display ads.

There could be other applications too. For instance, there has been interest recently[15] on how social media websites such as Facebook impact the news that people read, especially through algorithmic recommendations such as those for "Trending news". One way to estimate the impact of Facebook on news consumption would be to apply the Split-door criterion to browsing logs for articles from a news website. Here $Y_R$ would correspond to the visits that are referred from Facebook, and $Y_D$ would be all other direct visits to the news article. Whenever people's Facebook usage is not correlated with direct visits to a news article, we can identify the causal effect of Facebook's news suggestions.

Applications may not be restricted to the online world. For instance, Split-door criterion may be applied profitably to estimate the impact of discount coupons to sales in brick-and-mortar stores, as described in Section 3.

### 5.2 KEY ASSUMPTIONS

In the above applications though, we have to be careful to ensure that the Split-door criterion is applicable. The Split-door method allows us to test its applicability solely from observational data, by checking for independence between variables.

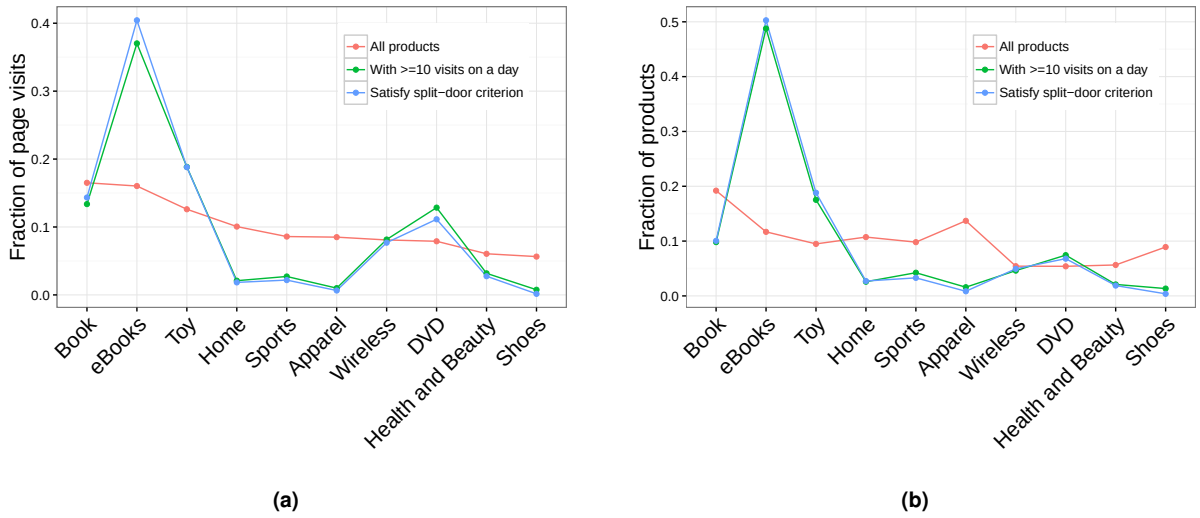



**Figure 9.** Distribution of products and total visits over product groups. Subset of products and page visits due to the focal products that satisfy the Split-door criterion are near identical to the set of the products with at least 10 visits on any day.

Compared to current methods[8] for justifying validity of natural experiments—justifications based on untestable assumptions—it offers a more precise specification of its validity.

Still, the statistical test for validating Split-door rests on two assumptions (*Connectedness* and *Faithfulness* assumptions). In general, for valid identification, $Y_D$ should be similar to $Y_R$ as possible. Ideally, they should be measures of the same outcome, e.g. visits to a product. In most practical cases, $Y_D$ and $Y_R$ are components of the same variable, so the connectedness assumption is reasonable. Further, barring coincidental equality of parameters, the faithfulness assumption is also unlikely to be violated.

However, in some (unlikely) cases, $U_{Y_r}$ may not be connected to $Y_D$ at all. In a recommender system, this can happen when the population of customers that browse product $j$ directly is completely different from users that visit product $i$. In such a scenario, we cannot say much without additional assumptions, because identification rests on the counterfactual behavior model of the users who visit $i$: whether they would search and be able to find $j$ through other channels in case no recommendations are shown.

In summary, the test for split-door validity does require two assumptions to hold, but because they do not proclaim about independence of variables, rather simply dependence between them, they are weaker assumptions than the current approaches to validating a natural experiment.

## 6 CONCLUSION

We presented a method for automatically finding natural experiments in observational data and showed its practical application in estimating the causal impact of a recommendation system. This method allows us to sample a bigger subset of the data for causal estimation, and also provides testable assumptions that can be verified from observational data. Going forward, we expect such data-driven approach for causal inference to become important, especially as access to more fine-grained data become available.

## References


1. Card, D. Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica* **69**, 1127–1160 (2001).
2. Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine* **27**, 1133–1163 (2008).
3. Sharma, A., Hofman, J. M. & Watts, D. J. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 453–470 (2015).
4. Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J. & Schölkopf, B. Distinguishing cause from effect using observational data: methods and benchmarks. *arXiv preprint arXiv:1412.3773* (2014).

5. Imbens, G. W. & Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press, 2015).
6. Pearl, J. *Causality* (Cambridge University Press, 2009).
7. Carmi, E., Oestreicher-Singer, G. & Sundararajan, A. Is oprah contagious? identifying demand spillovers in online networks. *NET Institute Working Paper* (2012).
8. Angrist, J. D. & Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion* (Princeton University Press, 2008).
9. Rosenzweig, M. R. & Wolpin, K. I. Natural" natural experiments" in economics. *Journal of Economic Literature* **38**, 827–874 (2000).
10. Morgan, S. L. & Winship, C. *Counterfactuals and causal inference* (Cambridge University Press, 2014).
11. Dunning, T. *Natural experiments in the social sciences: a design-based approach* (Cambridge University Press, 2012).
12. Spirtes, P., Glymour, C. N. & Scheines, R. *Causation, prediction, and search* (MIT Press, 2000).
13. Ricci, F., Rokach, L. & Shapira, B. *Introduction to recommender systems handbook* (Springer, 2011).
14. Agresti, A. A survey of exact inference for contingency tables. *Statistical Science* 131–153 (1992).
15. Flaxman, S., Goel, S. & Rao, J. M. Ideological segregation and the effects of social media on news consumption. *Available at SSRN 2363701* (2013).