Submitted to the Annals of Applied Statistics
arXiv: arXiv:1611.09414

SPLIT-DOOR CRITERION: IDENTIFICATION OF CAUSAL EFFECTS THROUGH AUXILIARY OUTCOMES

By Amit Sharma*, Jake Hofman* and Duncan Watts*

*Microsoft Research** †

We present a method for estimating causal effects in time-series data when fine-grained information about the outcome is available. Specifically, we examine what we call the *split-door* setting, when the outcome variable can be split up into two parts: one that is potentially affected by the cause, and another that is independent of it. 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. Using this independence test, we present a method that automatically finds subsets of the data where the split-door conditions are satisfied and computes the causal effect for this set. We demonstrate the method by estimating the causal impact of Amazon's recommender system, finding thousands of examples within the dataset that satisfy the split-door criterion, covering a quarter of available products in our dataset. Unlike past studies based on natural experiments that were limited to a single product category, we find eligible products across all popular categories with a near-identical distribution to the overall distribution of products. Further, we find that the widely-used click-through rate (CTR) metric overestimates the causal impact of recommender systems; across all product categories, at least half of the traffic attributed to recommender systems would have happened even without any recommendations. We also discuss other online and offline contexts where the split-door criterion can be applied, and provide an R package that implements the method.

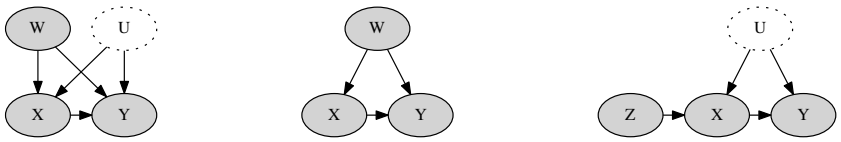
1. Introduction. The recent growth of digital platforms has generated an avalanche of highly granular and often longitudinal data regarding individual and collective behavior in a variety of domains of interest to researchers, including in e-commerce, healthcare, and social media consumption. Because the vast majority of this data is generated in non-experimental settings, researchers typically must deal with the possibility that any causal effects of interest are complicated by a number of potential confounds. For example, even effects as conceptually simple as the causal impact of recommendations on customer purchases are likely confounded by selection effects (Lewis, Rao and Reiley, 2011), correlated demand (Sharma, Hofman

Keywords and phrases: causal inference, data mining, causal graphical models, recommendation systems

and Watts, 2015), or other shared causes of both exposure and purchase. Figure 1a shows this canonical class of causal inference problems in the form of a causal graphical model (Pearl, 2009), where X is the cause and Y is its effect. Together U and W refer to all of the common causes of X and Y that may confound estimation of the causal effect, where critically some of these confounders (labeled W) may be observed, while others (U) are unobserved or even unknown. Ideally one would answer such questions by running randomized experiments on these platforms, but in practice such tests are possible only for the owners of the platform in question, and even then are often beset with implementation difficulties or ethical concerns (Fiske and Hauser, 2014). As a result researchers are left with two main strategies for making causal estimates from large-scale observational data, each with its own assumptions and limitations: either conditioning on observables or exploiting natural experiments.

1.1. *Background: Back-door criterion and natural experiments.* The first and by far the more common approach is to assume that the effect of unobserved confounders (U) is negligible after conditioning on the observed variables (W). Under such a *selection on observables* assumption (Imbens and Rubin, 2015), one can condition on W to estimate the effect of X on Y when these confounders are held constant. In the language of graphical models, this strategy is referred to as the *back-door criterion* (Pearl, 2009) on the grounds that the "back-door pathway" from X to Y (via W) is blocked by conditioning on W (see Figure 1b) and can be implemented by a variety of methods, including regression, stratification, and matching (Rubin, 2006; Stuart, 2010). Unfortunately for most practical problems it is difficult to justify that all of the important confounders have been observed. For example, consider the problem of estimating causal impact of a recommender system on e-commerce websites such as Amazon, where X would be the number of visits to a product's webpage, and Y the visits to a recommended product shown on that webpage. One could compute the observed click-through rate after conditioning on all user and product attributes on the grounds that these attributes provide a relative proxy for latent demand. Unfortunately, there are many unobserved confounders (e.g., advertising, media coverage, seasonality, etc.) that impact both a product and its recommendations, rendering the back-door criterion ineffective.

Motivated by the limitations of the back-door strategy, a second main approach is to identify an external event that affects the treatment X in a way that is arguably random with respect to potential confounds. The hope is that such variation, known as a *natural experiment* (Dunning, 2012), can



(a) Canonical causal inference problem

(b) Estimation with *back-door* criterion

(c) Estimation with Z as an *instrument*

Fig 1: *Left:* Graphical model for the canonical problem in causal inference and common estimation methods. We wish to estimate the effect of X on Y. W represents observed common causes of X and Y; U represents other unobserved (and unknown) common causes that confound observational estimates. *Middle*: The causal model under the *selection on observables* assumption, where there are no known unobserved confounds U. *Right:* The canonical causal model for an instrumental variable Z that systematically shifts the distribution of the cause X independently of confounds U.

serve as a proxy for an actual randomized experiment. Continuing with the problem of estimating the causal impact of recommendations, one might look for a natural experiment in which some products experience large and sudden changes in traffic, for instance when such a book is featured on Oprah's book club (Carmi, Oestreicher-Singer and Sundararajan, 2012). Assuming that the increase in traffic for a book is independent of demand for its recommendations, one can estimate the causal effect of the recommender by measuring the change in sales to the recommended products before and after the shock, arguing that these sales would not have happened in the absence of the recommender. Such shocks serve as *instrumental variables* that identify the effect of interest by shifting the distribution of the cause X independently of unobserved confounds U (Angrist, Imbens and Rubin, 1996). Figure 1c depicts this in a graphical model, where the additional observed variable Z denotes the instrumental variable.

These two main approaches trade off critical goals of identification and generalization in causal inference. Back-door conditioning can be applied to all available data, but provides no identification guarantees in the presence of unoberved confounders. Instrumental variables, in contrast, provide strong guarantees even in the presence of unobserved confounders, but these guarantees apply only in the relatively rare instances where a valid instrument that exogenously varies the cause X is available (e.g., lotteries (Angrist, Im-

bens and Rubin, 1996), variation in weather (Phan and Airoldi, 2015), or sudden, large events (Rosenzweig and Wolpin, 2000; Dunning, 2012)).

1.2. *The "split-door" criterion.* In this paper we introduce a causal identification strategy that incorporates elements of both the back-door and natural experiment approaches, but that applies in a different setting. Rather than conditioning on observable confounds W or exploiting sources of independent variation in the cause X, we instead look to *auxiliary outcomes* (Mealli and Pacini, 2013) to identify subsets of the data where causal identification is possible. Specifically, our strategy applies when the outcome variable Y can be effectively "split" into two constituents: one that is caused by X and the other that is independent of it. Figure 2a shows the corresponding causal graphical model, where Y_R denotes the "referred" outcome of interest affected by X and Y_D indicates the "direct" constituent of Y that does not directly depend on X. Returning to the recommender system example, Y_R corresponds to recommendation click-throughs on a product whereas Y_D would be all other traffic to that product that comes through channels such as direct search or browsing. Whenever such fine-grained data on Y is available, we show that it is possible to reduce causal identification to an independence test between the cause X and the auxiliary outcome Y_D. Because this strategy depends on the availability of a split set of variables for Y, we call it the *split-door* criterion for causal identification, by analogy with the more familiar back-door criterion.

While we make no assumptions about the functional form of relationships between unobserved and observed variables, a crucial assumption underlying the split-door criterion is that the auxiliary outcome Y_D must be affected by *all* unobserved common causes that confound the effect of X on Y_R. Although this assumption may appear strict, it becomes plausible in scenarios with fine-grained and accurate measurement of outcome channels, such as in digital systems. In the case of a recommender system, for example, this means that the number of direct and recommended visits to a product at any point in time are impacted (possibly differently) by the same components of demand for that product. This is the case whenever we can assume that people who visit the same product through recommendations or other channels are exchangeable, or less stringently, given any component of demand for a product, people with that demand are able to find the product through other channels such as search; they do not exclusively find the product through recommendations.

Under the above assumption, the split-door method seeks to identify subsets of the data where causal identification is possible. In this sense, the

(a) General split-door model: Outcome Y is split into Y_R and Y_D

(b) Valid split-door model: Data subsets where X is independent of U_Y

Fig 2: Panel (a) illustrates the canonical causal inference problem when outcome Y can be split up into two components. For clarity, unobserved confounders U are broken into U_Y that affects both X and Y, and U_X that affects only X. Split-door criterion finds subsets of the data where the cause X is independent of U_Y by testing independence of X and Y_D, leading to the unconfounded causal model shown in Panel (b).

method works analogously to a natural experiment. However, instead of looking for an exogenous instrument for a subset of the data that creates this variation in X, we look for variations in X directly. As in a natural experiment, it is important that any such variation in X is independent of potential confounds. For instance in the example above, it is important that a sudden burst of interest in a particular book is not correlated with changes in latent demand for its recommendations. To verify this, the split-door criterion relies on a statistical test to select for cases where there are no confounds (observed or otherwise) between X and Y_R. Specifically, we show that given a suitable auxiliary outcome Y_D and a test to establish if X and Y_D are independent, the causal effect between X and Y_R can be identified. Furthermore, since this test involves two observed quantities (X and Y_D), we can systematically search for subsets of the data that satisfy the required condition, potentially discovering a large number of cases in which we can identify the causal effect of X on Y_R.

We illustrate the utility of this method with a detailed example in which we estimate the causal impact of Amazon.com's recommendation system using historical web browsing data. Under the above assumption on the dependence between referred and direct visits to a product's webpage, we show how the criterion provides a principled mechanism for determining which subsets of the data to include in the analysis. The split-door criterion identifies thousands of such instances in a nine month period, comparable

in magnitude to a manually tuned approach using the same data (Sharma, Hofman and Watts, 2015), and an order of magnitude more than traditional approaches (Carmi, Oestreicher-Singer and Sundararajan, 2012). Further, the products included in our analysis are representative of the overall product distribution over product categories on Amazon.com, thereby improving both the precision and generalizability of estimates. Consistent with previous work (Sharma, Hofman and Watts, 2015), we find that observational estimates of recommendation click-through rates (CTRs) overstate the actual effect by at least two-fold, thus questioning the applicability of popular CTR metrics for assessing the impact of recommendation systems. For application to other online and offline scenarios, we provide an R package[1] that implements the split-door criterion.

1.3. *Outline of paper.* The remainder of this paper proceeds as follows. In section 2 we start with a formal definition of the split-door criterion and and give precise conditions under which the split-door criterion holds. For clarity we provide proofs for causal identification both in terms of the causal graphical model above and also in terms of structural equations. In section 2.3 we propose a simple, scalable algorithm for identifying causal effects using the split-door criterion. Then in section 2.4, we explain more formally how the split-door criterion differs from the instrumental variables and back-door methods mentioned above. Section 3 presents details about the Amazon.com data and application of the split-door criterion to estimate the causal impact of its recommendation system. In section 4 we then discuss limitations of the split-door criterion as well as other settings in which the criterion applies, arguing that many existing datasets across a variety of domains have the structure that outcomes of interest can be decomposed into their "direct" and "referred" constituents. Finally, in section 5 we conclude with a prediction that as the size and granularity of available datasets, and the number of variables in them, increase at an ever faster rate, data-driven approaches to causal identification will become commonplace.

2. The Split-door Identification Criterion. The split-door criterion can be used whenever observed data is generated from the model shown in Figure 2a. Here X represents the cause of interest, Y_R denotes the "referred" portion of the outcome affected by it, and Y_D indicates the "direct" part of the outcome which does not directly depend on X. We denote the overall outcome by $Y = Y_R + Y_D$. We let U_Y represent all unobserved causes of Y, some of which may also be common causes of X, hence the arrow

[1]URL: http://www.github.com/amit-sharma/splitdoor-causal-criterion

from U_Y to X. Additional latent factors that affect only X are captured by U_X. Both U_X and U_Y can be a combination of many variables, some observed and some unobserved. (For full generality, the analysis presented here assumes that all confounds are unobserved.) The unobserved variables U_Y make it difficult to estimate the causal effect of X on Y; that is, they create "back-door pathways" that confound the relationship of interest, resulting in biased estimates of the causal effect. The central idea behind the split-door criterion is that we can use an independence test between the auxiliary outcome Y_D and X to systematically search for subsets of the data that are free of these confounds and do not contain back-door pathways between X and Y_R. In other words, we can conclude that such subsets of the data were generated from the unconfounded causal model shown in Figure 2b, and therefore the causal effect of X on Y can be estimated directly from these data. Importantly, identification of the causal effect rests on the assumption that no part of U_Y causes one part of Y and not the other.

2.1. *The split-door criterion through a graphical model.* Here we formalize the intuition above in the causal graphical model framework. To identify the causal effect, we make the following two assumptions. The first pertains to connectedness of the causal model.

ASSUMPTION 1 (Connectedness). *Any unobserved confounder U_Y that causes Y_R also causes Y_D. Therefore, the causal effect of such U_Y on Y_D is non-zero.*

Note that the connectedness assumption requires only that the causal effect of U_Y on Y_D be non-zero, with no other requirements on the size of the effect(s) involved. That said, it is a strong requirement in general, as it applies to all sub-components of U_Y and thus involves assumptions about potentially high-dimensional, unobserved variables. That is, *Assumption 1* implies that any unobserved sub-component U_{Yi} of U_Y that causes Y_R also causes Y_D and the causal effect of such U_{Yi} on Y_D cannot be zero. Whenever Y_D and Y_R are components of the same variable, it is plausible that they share causes, but one still must establish that this condition holds to ensure causal identification. It is instructive to compare this assumption to strict independence assumptions involving unobserved confounders required by methods such as instrumental variables (Angrist, Imbens and Rubin, 1996), which are also difficult to justify for observational data in general.

In addition, we make the standard assumption connecting statistical and causal independence between observed variables (related to *Faithfulness* (Spirtes, Glymour and Scheines, 2000) or *Stability* (Pearl, 2009)). This as-

sumption serves to rule out an event where incidental equality of parameters or certain data distributions render two variables statistically independent even though they are causally related. This is a general requirement for causal discovery from observational data for this and many other methods.

ASSUMPTION 2 (Independence). *If any two observed variables are statistically independent, then they are also causally independent in the graphical model of Figure 2a.*

Specifically, for the model shown in Figure 2a, the above assumption rules out the possibility of an (unlikely) event where the effect of the unobserved confounders U_Y cancels out exactly in a way such that X and Y_D become statistically independent.

While the above two assumptions are conceptually simple and are sufficient for proving the validity of the split-door criterion, there are in fact weaker versions of each that remain sufficient for establishing causal identification.

Assumption 1a. *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.*

That is, it is not required that Y_D be affected by all U_Y that cause Y_R, only those sub-components of U_Y that also cause X.

Assumption 2a. *If X and Y_D are statistically independent, then they are also causally independent in the graphical model of Figure 2a.*

That is, we simply require the Independence assumption to hold for X and Y_D.

Under Assumptions 1a and 2a, we can show that statistical independence of X and Y_D ensures that X is not confounded by U_Y. First, we provide a result about the resulting causal graph structure when $X \perp\!\!\!\perp Y_D$.

LEMMA 1. *Let X, Y_R and Y_D be three observed variables corresponding to the causal model in Figure 2a, where U_Y refers to unobserved causes of Y_R. If the Connectedness (1a) and Independence (2a) assumptions hold, then $X \perp\!\!\!\perp Y_D$ implies that the edge $U_Y \to X$ does not exist or that U_Y is constant.*

Proof (Argument). The proof can be completed directly from Figure 2a and properties of a causal graphical model.

$X \perp\!\!\!\perp Y_D$ implies that the causal effect of U_Y on Y_D and X somehow cancels out on the path $X \leftarrow U_Y \to Y_D$. By *Assumption 2a*, this cancellation

is not due to incidental equality of parameters or a particular data distribution, but rather a property of the causal graphical model. Therefore, this can only happen if
(i) U_Y is constant (and thus *blocks* the path), or
(ii) One of the edges exists trivially (does not have a causal effect). Using *Assumption 1a*, U_Y has a non-zero effect on Y_D. Then, the only alternative is that $X \leftarrow U_Y$ edge does not exist, leading to the unconfounded causal model in Figure 2b.

PROOF. We provide a proof by contradiction using the principle of *d-separation* Pearl (2009) in a causal graphical model.

Let us suppose $X \perp\!\!\!\perp Y_D$, and that $U_Y \rightarrow X$ edge exists and U_Y is not constant.

Using the rules of *d-separation* on the causal model in Figure 2a, the path X-U_Y-Y_D corresponds to:

$$(X \perp\!\!\!\perp Y_D | U_Y)_G \tag{2.1}$$

$$(X \not\perp\!\!\!\perp Y_D)_G \tag{2.2}$$

where the notation $(.)_G$ refers to *d-separation* under a causal model G. In our case, G corresponds to the causal model in Figure 2a.

However, using Assumption 2a, statistical independence of X and Y_D implies causal independence, and thus, d-separation of X and Y_D.

$$(X \perp\!\!\!\perp Y_D)_G \tag{2.3}$$

Equations 2.2 and 2.3 result in a contradiction. To resolve,
(i) Either U_Y is constant and thus 2.1 implies $(X \perp\!\!\!\perp Y_D)_G$ holds, or
(ii) Path X-U_Y-Y_D does not exist. Using *Assumption 1a* of dependence of Y_D on U_Y, the only possibility is that $X \leftarrow U_Y$ edge does not exist.

□

We now show that Lemma 1 removes confounding due to U_Y and that the observational estimate $P(Y_R|X = x)$ is also the causal estimate.

THEOREM 2.1 (Split-door Criterion). *Under the assumptions of Lemma 1, the causal effect of X on Y_R is not confounded by U_Y and is given by:*

$$P(Y_R|do(X = x)) = P(Y_R|X = x)$$

where $do(X = x)$ refers to experimental manipulation of X and $Y_R|X = x$ refers to the observed conditional distribution.

Proof (Argument). Lemma 1 leads to two cases:
(i) By the back-door criterion (Pearl, 2009), if U_Y is constant, then X and Y_R are unconfounded, because the only back-door path between X and Y_R contains U_Y on it.
(ii) Similarly, if $U_Y \to X$ edge does not exist, then X and Y_R are unconfounded because the absence of $U_Y \to X$ edge removes the back-door path between X and Y_R.

In both (i) and (ii) cases, unconfoundedness implies that the effect of X on Y_R can be estimated by the observational distribution.

PROOF. The proof follows from an application of the second rule of do-calculus (Pearl, 2009).

$$P(\mathcal{Y}|do(\mathcal{Z}=z),\mathcal{W}) = P(\mathcal{Y}|\mathcal{Z}=z,\mathcal{W}) \quad \text{if} \quad (\mathcal{Y} \perp\!\!\!\perp \mathcal{Z}|\mathcal{W})_{G_{\underline{\mathcal{Z}}}} \tag{2.4}$$

where $G_{\underline{Z}}$ refers to the underlying causal graphical model with all outgoing edges from $\mathcal{Z}$ removed.

Substituting $\mathcal{Y} = Y_R$, $\mathcal{Z} = X$, $G_{\underline{X}}$ corresponds to the causal model from Figure 2a without the $X \to Y_R$ edge. Using Lemma 1, two cases exist:

(i) U_Y is constant
Let $\mathcal{W} = U_Y$. Under the modified causal model $G_{\underline{X}}$ without $X \to Y_R$ edge, the path X-U_Y-Y_R is the only path connecting X and Y_R, which leads to the following *d-separation* result:

$$(Y_R \perp\!\!\!\perp X|U_Y)_{G_{\underline{X}}} \tag{2.5}$$

Combining Rule 2.4 and the above *d-separation* result, we obtain

$$P(Y_R|do(X=x),U_Y) = P(Y_R|X=x,U_Y) = P(Y_R|X=x)$$

where the last equality holds because U_Y is constant throughout.

(ii) Edge $U_Y \to X$ does not exist.
Let $\mathcal{W} = \emptyset$. Under the modified causal model $G_{\underline{X}}$ without $X \to Y_R$ edge, X and Y_R are trivially *d-separated* because no path connects them without the edge $U_Y \to X$.

$$(Y_R \perp\!\!\!\perp X)_{G_{\underline{X}}} \tag{2.6}$$

From Rule 2.4 and the above *d-separation* result, we obtain

$$P(Y_R|do(X=x)) = P(Y_R|X=x)$$

□

2.2. *The split-door criterion through structural equations.* Although we have motivated the split-door criterion in terms of the causal model in Figure 2a, for expositional clarity we note that it is also possible to express it in terms of structural equations. Specifically, we can write three structural equations:

$$x = g(u_x, u_y, \varepsilon_x) \qquad y_r = f(x, u_y, \varepsilon_{yr}) \qquad y_d = h(u_y, \varepsilon_{yd}), \tag{2.7}$$

where ε_x, ε_{yr}, and ε_{yd} are mutually independent, zero-mean random variables that capture modeling error and statistical variability. As in *Assumption 1a*, we assume that U_Y affects both Y_D and Y_R. In general, the causal effects among variables may not be linear; however, for the purpose of building intuition we rewrite the above equations in linear parametric form:

$$x = \eta u_x + \gamma_1 u_y + \epsilon_x \qquad y_r = \rho x + \gamma_2 u_y + \epsilon_{yr} \qquad y_d = \gamma_3 u_y + \epsilon_{yd}, \tag{2.8}$$

where ρ is the causal parameter of interest, and ϵ_x, ϵ_{yr} ϵ_{yd} are independent errors in the regression equations. The split-door criterion requires independence of X and Y_D, which in turn implies that $Cov(X, Y_D) = 0$:

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

Assuming that Y_D is affected by U_Y (and therefore γ_3 is not 0), the above can be zero only if $\gamma_1 = 0$, or if U_Y is constant ($\mathrm{Var}[U_Y] = 0$). 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:

$$y_r = \rho x + \epsilon'_{yr} \tag{2.9}$$

where ϵ'_{yr} is now independent, zero-mean noise.

2.3. *Algorithm for Applying the Split-door Criterion.* The above results point to an algorithm for applying the split-door criterion to observational data. Specifically, given an empirical test for independence between the cause X and the auxiliary outcome Y_D, we can select instances in our data that pass this test and satisfy the split-door criterion. In this section we develop such a test for time series data, resulting in a simple, scalable identification algorithm.

At a high level, the algorithm works as follows. First, divide the data into equally-spaced time intervals τ such that each interval has enough data points to reliably estimate the joint probability distribution $P(X, Y_D)$. Then, for each time interval τ,

1. Determine whether X and Y_D are independent using an empirical independence test.
2. If $X \perp\!\!\!\perp Y_D$, then use the observed conditional probability $P(Y_R|X = x)$ to estimate the causal effect in the interval τ. Otherwise exclude the interval τ from the analysis.
3. Average over all time intervals where $X \perp\!\!\!\perp Y_D$ to obtain the mean causal effect of X on Y_R.

Implementing the algorithm requires deciding suitable choices for an independence test and its significance level, taking into account multiple comparisons.

2.3.1. *Choosing an independence test.* Each X-Y_D pair in *Step 1* provides two vectors of size τ with observed values for X and Y_D. The key decision is whether these data vectors are independent of each other. In theory any empirical test that reliably establishes independence between X and Y_D is sufficient to identify instances where the split-door criterion applies. For instance, assuming we have enough data, we could test for independence by comparing the empirical mutual information to zero (Steuer et al., 2002; Pethel and Hahs, 2014). In practice, however, because we consider subsets of the data over relatively small time intervals τ, there may be substantial limits to the statistical power we have in testing for independence. For example, it is well known that in small sample sizes, testing for independence via mutual information estimation can be heavily biased (Paninski, 2003).

Thus, when working with small time intervals τ we recommend the use of exact independence tests and randomization inference (Agresti, 1992, 2001; Lydersen et al., 2007). When X and Y_D are discrete variables, methods such as Fisher's exact test are appropriate. If, however, X and Y_D are continuous—as is this case for the example we study in Section 3—we recommend the use of resampling-based randomization inference for establishing independence. In general, this involves repeatedly sampling randomized versions of the empirical data to simulate the null hypothesis and then comparing a test statistic on the observed data to the same on the null distribution. Specifically, we recommend the following. For each X-Y_D pair, simulate the null hypothesis of independence between X and Y_D by replacing the observed X vector with a randomly sampled vector from the overall empirical distribution of X values. From this simulated X-Y_D instance, compute a test statistic that captures statistical dependence, such as the distance correlation, which is able to detect both non-linear and linear dependence (Székely et al., 2007; de Siqueira Santos et al., 2014). Repeat this multiple times to obtain a null distribution for the test statistic of this X-Y_D pair. Finally,

compute the probability p of obtaining a test statistic as extreme as the observed statistic under the null distribution, and select instances in which the probability p is above a pre-chosen significance level α.[2]

2.3.2. *Choosing a significance level.* Deciding a pre-chosen significance level, however, is non-trivial. In general, higher p-values are desirable, since our goal is to identify the subsets where X and Y_D are independent, rather than refute their independence. In other words, we are interested in a low Type II error (or false negatives), in contrast to standard null hypothesis testing, where the focus is on Type I errors (false positives) and hence significance levels are set low. Therefore, one way to choose a significance level would be to choose α as close as possible to 1 to minimize Type II errors when X and Y_D are dependent. At the same time, we need to ensure that the test yields adequate power for finding independent $X - Y_D$ pairs. Unlike a conventional hypothesis test for dependent pairs, power for our test is the probability that the test declares an $X - Y_D$ pair to be independent given that it is actually independent. This is given by $1 - \alpha$. As we increase α, type II errors decrease, but power also decreases.

Moreover, the combination of low power and a large number of hypothesis tests raises concerns about falsely accepting pairs that are actually dependent. As an extreme example, even when all X-Y_D pairs in a given dataset are dependent, some of them will pass the independence test simply due to random chance. Therefore, a more principled approach to selecting α comes through estimating the expected fraction of erroneous split-door instances returned by the procedure, which we refer to as ϕ. We apply techniques from the multiple comparisons literature (Storey, 2002; Liang and Nettleton, 2012; Farcomeni, 2008) and provide a method to estimate this fraction for any given significance level. Details are in Appendix A.

2.4. *Connections to other methods.* The split-door criterion is an example of methods that provide data-driven identification of causal effects under certain assumptions (Jensen et al., 2008; Sharma, Hofman and Watts, 2015; Grosse-Wentrup et al., 2016; Cattaneo, Frandsen and Titiunik, 2015). By searching for subsets of the data where desired independence holds, it also shares some properties with natural experiment methods such as instrumental variables and conditioning methods such as regression. We discuss these connections below; table 3 provides a summary for easy comparison.

[2]In contrast to standard hypothesis testing where one is looking to reject the null hypothesis that two variables are independent and therefore thresholds on a small p-value, here we are looking for independent X-Y_D pairs that are highly probable under the null and thus want a large p-value.

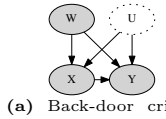
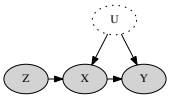
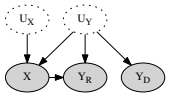
Graphical model	Description	Untestable assumptions	Limitations	Recommendations example
 **(a)** Back-door criterion	Condition on observed confounders W to isolate treatment effect.	$X \perp\!\!\!\perp U$ or $Y \perp\!\!\!\perp U$	Unlikely that there are no unobserved confounders U.	Regress click-throughs on product attributes and direct visits to recommended product.
 **(b)** Instrumental variable	Analyze subset of data having independent variation in the treatment.	$Z \perp\!\!\!\perp U$ and $Z \perp\!\!\!\perp Y \mid X, U$	Hard to find a source of exogenous variation in treatment.	Measure marginal click-throughs on subset of products that experience large, sudden shocks in traffic.
 **(c)** Split-door criterion	Analyze subset of data where auxiliary outcome Y_D is independent of treatment.	$Y_D \not\perp\!\!\!\perp U_Y$	Requires dependency between an auxiliary outcome and all confounders.	Measure marginal click-throughs on all pairs of products that have uncorrelated direct traffic.

FIG 3

Comparison of methods for the canonical problem in causal inference: estimating the effect of treatment X on outcome Y. U represents all the unobserved (and unknown) confounders that commonly cause both X and Y.

2.4.1. *Instrumental Variables.* Both the split-door criterion and instrumental variable (IV) methods can be used to exploit naturally occurring variation in subsets of observational data to identify causal effects. Importantly, however, they make different assumptions. In IV methods, one uses an auxiliary variable Z, called an instrument, that is assumed to be exogenous and that systematically shifts the distribution of the cause X. The validity of an instrument relies on two additional assumptions: first that it is effectively random with regard to potential confounders ($Z \perp\!\!\!\perp U$), and second that the instrument affects the outcome Y only through the cause X ($Z \perp\!\!\!\perp Y|X, U$). Both of these conditions involve independence claims between observed and unobserved variables, making them impossible to test in practice (Dunning, 2012).

The split-door criterion also relies on an auxiliary variable, but in a different setting. Specifically, it exploits an auxiliary outcome Y_D that serves as a proxy for unobserved common causes U_Y under three important assumptions. The first is that the cause X does not affect Y_D directly. The second assumption requires that all unobserved confounders (between the cause and outcome) that affect Y_R also affect Y_D. As with IV methods above, these two assumptions involve knowledge of an unobserved variable and, as a result, cannot be tested. The third assumption requires independence between the cause X and the auxiliary outcome Y_D. Since both of these variables are observed, this assumption can be tested empirically so long as we are in the

standard setting where statistical independence implies causal independence (*Assumption 2a*), equivalent to the assumption of *faithfulness* (Spirtes, Glymour and Scheines, 2000). This rules out unlikely cases where we obtain a particular distribution where statistical independence happens by chance.[3]

It is difficult to compare these two sets of assumptions in general, but in different scenarios, one of these methods may be more suitable than the other. If a valid instrument is known to exist, for instance through changes in weather or as a result of a lottery, the variation it produces can and should be exploited to identify causal effects of interest. The split-door criterion, in contrast, is most useful when one suspects there is random variation in the data, but cannot identify its source *a priori*. In particular, it is well-suited for large-scale data where the first two assumptions are plausible, such as in digital or online systems. Further, even if the split-door assumptions are not fully satisfied, the criterion can nonetheless be used to find time intervals with a potentially exogenous variation. Such time intervals can be manually inspected to check if there are known sources of exogenous variation (such as a large and sudden external shock) corresponding to the natural variation observed. This can be useful in IV methods because it provides a filtered view of observed data, wherein some variations in the cause may be manually traced to actual exogenous events. As an example, recent work on a data-driven search for instrumental variables bridges the gap between the two methods by algorithmically searching for a specific kind of instrument: a large and sudden shock to time series data (Sharma, Hofman and Watts, 2015). The split-door criterion can be viewed as an extension of this approach.

2.4.2. *Methods based on empirical independence tests.* Similar to the approach taken here, recent work proposes a data-driven method for determining the appropriate window size in regression discontinuity designs (Cattaneo, Frandsen and Titiunik, 2015; Cattaneo, Titiunik and Vazquez-Bare, 2017). In regression discontinuities, treatment (e.g., acceptance into a program) is assigned based on whether an observed variable (e.g., a test score) is above or below a pre-determined cutoff. The assumption is that one can compare outcomes for those just above and just below the cutoff to estimate causal effects, but the central problem is how far from the cutoff this assumption holds. The authors present a data-driven approach for selecting a window by testing for independence between the treatment and pre-

[3]This might be a concern in an adversarial setting (e.g. when Nature or system designers have incentive to invalidate the method), however, this is not the case in most observational data.

determined covariates that are uncoupled to the outcome of interest. This is similar to the split-door approach in that both use independence tests to determine which subsets of the data to include when making a causal estimate. As a result, both methods are subject to concerns around multiple hypothesis testing, although the regression discontinuity setting typically involves many fewer comparisons than the split-door method (dozens instead of the thousands we analyze here) and occurs over nested windows. For these reasons we treat multiple comparisons differently, estimating the error rate in identifying independent instances instead of adjusting nominal thresholds to try to eliminate them.

2.4.3. *Back-door criterion.* Alternatively, the split-door criterion can be interpreted as using Y_D as a proxy for all confounders U_Y, and estimating the causal effect whenever Y_D (and hence U_Y) is independent of X. Such an approach may appear to be nothing more than a variant of the back-door criterion (Figure 1b) where one conditions on Y_D instead of U_Y, however there are two key differences between the two methods.

First, substituting Y_D for U_Y in the back-door criterion assumes that Y_D is a perfect proxy for U_Y. This is a much stronger assumption than requiring that Y_D is simply affected by U_Y, because any difference (measurement error) between Y_D and U_Y can invalidate the back-door criterion (Spirtes, Glymour and Scheines, 2000). Further, the assumption about Y_D being affected by U_Y becomes more plausible when Y_D is simply an additive component of Y. Second, the two methods differ in their approach to identification. The split-door criterion *controls* for the effect of unobserved confounders by finding subsets of data where X is not affected by U_Y, whereas the back-door criterion *conditions* on a proxy for U_Y to nullify the effect of unobserved confounders. Therefore, by directly controlling at the time of data selection, the split-door criterion focuses on admitting a subset of the data for analysis and simplifies effect estimation, whereas methods based on back-door criterion such as regression, matching, and stratification process the whole dataset and extract estimates via statistical models (Morgan and Winship, 2014).

To illustrate these differences, we compare mathematical forms of the split-door and back-door criteria in terms of regression equations. Conditioning on Y_D using regression will lead to the following equation

$$y_r = \rho'' x + \beta y_d + \epsilon''_{yr},$$

applied to the entire dataset. In contrast the split-door criterion leads to the

Fig 4: Screenshot of a focal product, the book "Purity", and its recommendations on Amazon.com

simpler equation (as shown earlier in Section 2.2)

$$y_r = \rho x + \epsilon'_{yr},$$

applied only to subsets of data where X and Y_D are independent.

3. Application: Impact of a Recommender System. We now apply the split-door criterion to the problem of estimating the causal impact of Amazon.com's recommender system. Recommender systems have become ubiquitous in online settings, providing suggestions for what to buy, watch, read or do next (Ricci, Rokach and Shapira, 2011). Figure 4 shows an example of one of the millions of product pages on Amazon.com, where the main item listed on the page, or *focal product*, is the book "Purity" by Jonathan Franzen. Listed alongside this item are a number of *recommended products*—two written by Franzen and one by another author—suggested by Amazon as potentially of interest to a user looking for "Purity". Generating and maintaining these recommendations takes considerable resources, and so a natural question one might ask is how exactly exposure to these recommended products changes consumer activity.

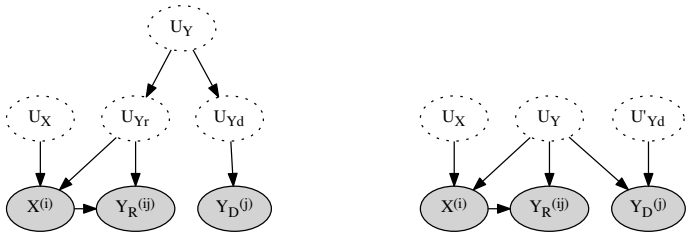
While simple to state, this question is difficult to answer because it requires an estimate of the counterfactual of what would have happened had someone visited a focal product but had not been exposed to any recommendations. Specifically, we would like to know how much traffic recommender systems *cause*, over and above what would have happened in their absence. Naively one could assume that users would not have viewed these other products without the recommender system, and as a result simply compute the observed click-through rate on recommendations (Mulpuru, 2006; Grau, 2009). But this ignores complications that arise due to correlated demand: users might have found their way to some of these recommended products

anyway via direct search or browsing, which we collectively refer to as "direct traffic". For instance, users who are interested in the book "Purity" might be fans of Franzen in general, and so might have *directly* searched on Amazon.com for some of his other works such as "Freedom" or "The Corrections", even if they had not been shown recommendations linking to them. The key to properly estimating the causal impact of the recommender, then, lies in accounting for correlated demand between a focal product and its recommendations. For instance, it might happen to be the case that people searching for "Purity" might not independently seek out the third recommendation shown, "City on Fire", in the absence of the recommender.

In this section we show how the split-door criterion can be used to eliminate this issue of correlated demand by automatically identifying and analyzing examples like this, where demand for a product and one (or more) of its recommendations are independent over some time interval τ. We do so by first formalizing this problem through a causal graphical model of recommender system traffic, revealing a structure amenable to the split-door criterion. Then we apply the criterion to a large-scale dataset of web browsing activity on Amazon.com to discover thousands of instances satisfying the criterion. Our results show that a naive observational estimate of the impact of this recommender system overstates the causal impact on the products analyzed by a factor of at least two. We conclude with a number of robustness checks and comments on the validity and generalizability of our results.

3.1. *Building the causal model.* The above discussion highlights that unobserved common demand for both a focal product and its recommendations can introduce bias in naive estimates of the causal click-through rate (CTR) on recommendations. We formalize this in the graphical model shown in Figure 5a, with the following variables defined and aggregated for each day:

- X denotes the number of visits to the focal product i's webpage.
- Y_R denotes the number of visits to the recommended product j, through clicks on the recommendation for product j on product i's webpage.
- Y_D is the number of *direct* visits to product j that did not occur through clicking on a recommendation. These could be visits to j from Amazon's search page, or through a direct visit to j's webpage URL.
- U_Y denotes the overall unobserved demand for product j, which manifests itself as unobserved demand U_{Y_r} for recommendation click-throughs for product j and unobserved demand U_{Y_d} for direct visits to product j.
- U_X represents the part of unobserved demand for product i that is



(a) (b)

Fig 5: Equivalent causal graphical models for estimating the impact of "similar product" recommendations on Amazon.com.

independent of U_{Y_r}.

We wish to estimate the causal effect of X on Y. As in the graphical model for the split-door we introduced in Figure 2a, Y is split up into two constituents: Y_R and Y_D, and X directly causes only one of them. We refer to Y_R as *recommendation* visits and Y_D as *direct* visits to the recommended product j.

At first glance, Figure 5a appears to differ from the the split-door model, because U_Y is separated into two components, U_{Y_R} and U_{Y_D}. This is because customers who visit product j through recommendation from product i are potentially different from those who visit j directly. In other words, we expect a different population of users to visit product j directly.[4] However, we can transform the causal graphical model to produce an equivalent model amenable to the split-door criterion. Specifically, we collapse components of U_Y, U_{Y_r} and U_{Y_d} to generate Figure 5b. Overall unobserved demand U_Y for product j now affects both Y_R and Y_D. Intuitively, the causal model in Figure 5b assumes 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, since both Y_R and Y_D correspond to visits to the same product j. Note that as in the structural equations in Section 2, each of the observed variables in the causal graphical model also has an independent error term associated with it, which we do not show in the graph.

[4]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.

To apply the split-door criterion, we must investigate the plausibility of the *Connectedness* and *Independence* assumptions from Section 2.1. As mentioned above, Y_R and Y_D are additive components of visits to the same product j, so it is reasonable to assume that they are both affected (possibly differently) by the same components of demand U_Y for the product j. To the extent that each component of the demand brings users to visit the product both through recommendation visits and direct visits (and not exclusively through recommendation visits), the *Connectedness* assumption for the product's demand is expected to hold. As for the *Independence* assumption, while we cannot rule out coincidental cancellation of effects that result in $X \perp\!\!\!\perp Y_D$ and violate the assumption, we expect such events to be unlikely over a large number of product pairs. Further, for complementary product recommendations (which are the focus of this paper), we can logically rule out violation of the *Independence* assumption because the demand for two complementary products are expected to be positively correlated with each other. Therefore, it is reasonable to assume that the unobserved demand U_Y (and all its sub-components) affect both X and Y_D in the same direction. For instance, let the effect of U_Y be increasing for both X and Y_D. Then the *Independence* assumption is satisfied because the effect of U_Y cannot be canceled out on the path $X \leftarrow U_Y \rightarrow Y_D$ if the effects of U_Y (and any of its sub-components) on X and Y_D are all positive. Given the above assumptions, the same reasoning from Section 2.1 allows us to derive $X \perp\!\!\!\perp Y_D$ as a sufficient condition for causal identification.

3.2. *Browsing data.* Estimating the causal impact of Amazon.com's recommender system requires fine-grained data detailing activity on the site. To obtain such information, we turn to anonymized browsing logs from users who have installed the Bing Toolbar and have consented to provide their browsing data through it. These logs cover a period of nine months from September 2013 to May 2014 and contain a session identifier, an anonymous user identifier, and a time-stamped sequence of all non-secure URLs that the user visited in that session. We restrict our attention to browsing sessions on Amazon.com, which leaves us with 23.4 million page views by 2.1 million users to 1.3 million unique products. Of these products, we examine those that receive a minimum of 10 page views on at least one day in this time period, resulting in roughly 22,000 focal products of interest.

Amazon shows many different kinds of recommendations on its site. We limit our attention here to the "Customers who bought this also bought" recommendations depicted in Figure 4, as these recommendations are the most common and are shown on product pages from all product categories.

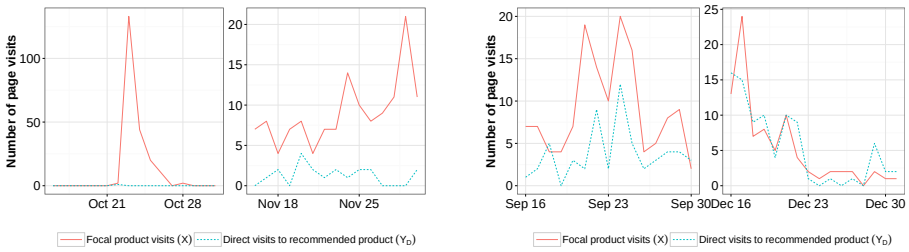
To apply the split-door criterion, we need to identify focal product and recommended product pairs from the log data and separate out traffic for recommended products into direct (Y_D) and recommended (Y_R) visits. Fortunately it happens to be the case that Amazon makes this possible by explicitly embedding this information in their URLs. Specifically, given a URL for an Amazon.com page view, we can use the `ref`, or referrer, parameter in the URL to determine if a user arrived at a page by clicking on a recommendation or by other means. We then use the sequence of page views in a session to identify focal and recommended product pairs by looking for focal product views that precede recommendation views. Further details about the toolbar dataset and construction of focal and recommended product pairs can be found in past work (Sharma, Hofman and Watts, 2015).

3.3. *Applying the split-door method.* Having argued for the split-door criterion and extracted the relevant data from browsing logs, the final step in estimating the causal effect of Amazon.com's recommendation system is to use the criterion to search for instances where a product and its recommendation have uncorrelated demand. Recalling Section 2.3, this amounts to employing a statistical test to identify instances where the direct traffic to a product and one (or more) of its recommendations is independent. As recommended there, we employ a randomization test and search for 15 day time periods that fail to reject the null hypothesis that these two time series were independently generated. The choice of $\tau = 15$ days is guided by empirical considerations: we require a time period over which we will have enough data for estimation, but also during which Amazon's recommendations are expected to stay constant.

The full application of the split-door criterion is as follows. For each focal product i and each $\tau = 15$ day time period:

1. Compute $X^{(i)}$, the number of visits to the focal product on each day, and $Y_R^{(ij)}$, the number of click-throughs to each recommended product j. Also record the total direct visits $Y_D^{(j)}$ to each recommended product j.
2. For each recommended product j, use the randomization test from Section 2.3 to determine if $X^{(i)}$ is independent of $Y_D^{(j)}$ at a pre-specified significance level.[5]
    - If $X^{(i)}$ is found to be independent of $Y_D^{(j)}$, compute the observed click-through rate (CTR) as $\hat{\rho}_{ij\tau} = (\sum_{t=1}^{\tau} Y_R^{(ij)})/(\sum_{t=1}^{\tau} X^{(i)})$ dur-

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



(a) Accepted at threshold=0.95 (b) Rejected at threshold=0.95

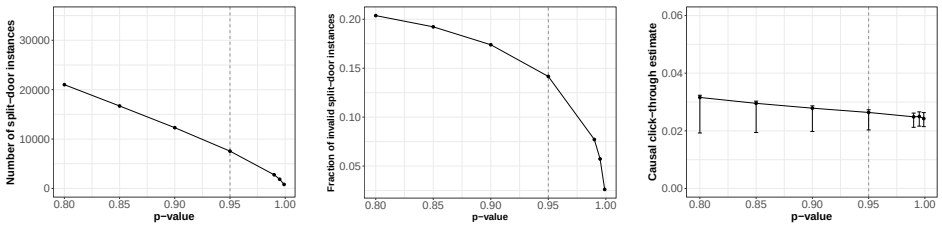
Fig 6: Examples of visit time series for focal and recommended products that are accepted or rejected by the split-door criterion for $\alpha = 0.95$.

ing this period as the causal estimate of the CTR. Otherwise ignore this product pair.

3. Aggregate the causal CTR estimate over all recommended products to compute the total causal CTR per focal product, $\hat{\rho}_{i\tau}$.

Finally, average the causal CTR estimate over all time periods and focal products to arrive at the mean causal effect, $\hat{\rho}$, and compute the rate of erroneous split-door instances ϕ to estimate error in this estimate, as detailed in the Appendix.

3.4. *Results.* Applying the above algorithm results in over 114,000 potential split-door instances, where each instance consists of a pair of focal and recommended product over a 15 day time period. Out of these, we obtain more than 7,000 instances that satisfy the split-door criterion at a significance level of $\alpha = 0.95$. Figure 6a shows examples of product pairs that are accepted by the test. The example on the left shows a focal product that receives a large a sudden shock in traffic, while direct traffic to its recommended product remains relatively flat. This is reminiscent of the examples analyzed in Carmi, Oestreicher-Singer and Sundararajan (2012) and Sharma, Hofman and Watts (2015). The example on the right, however, shows the more general patterns that are accepted under the split-door criterion but not considered by these previous approaches: although direct traffic to both the focal and recommended products vary substantially, they do so independently, and so are still useful in our estimate of the recommender's effect. Conversely, two example product pairs that are rejected by the test are shown in Figure 6b. As is visually apparent, each of the focal and recommended traffic patterns are highly correlated, and therefore not



(a) (b) (c)

Fig 7: Subplot (a) shows the number of valid split-door instances obtained as the p-value threshold (α) is increased. Subplot (b) shows the expected fraction of erroneous instances (ϕ) returned by the method for those values of α. The corresponding estimate for causal CTR is shown in Subplot (c); error bars account for both bias due to ϕ and natural variance in the mean estimate.

useful in our analysis.

The number of valid instances increases as we decrease the nominal p-value threshold, leading to over 20,000 split-door instances for $\alpha = 0.80$, covering nearly 11,000 unique focal products (Figure 7a). However, as detailed in Appendix A, the expected fraction of invalid instances ϕ increases to 0.21 (Figure 7b), indicating that one in five of the returned split-door instances may be invalid. As we increase the p-value threshold, ϕ decreases but so do the number of split-door instances returned by the method.

A better view of this tradeoff is obtained through Figure 7c, which shows the split-door estimate computed at different p-value thresholds, along with estimated error bars using Equation B.1 from Appendix B. These error bars account for both bias due to erroneous split-door instances and the natural variance in the mean estimate due to sampling.[6] At $\alpha = 0.999$, error due to ϕ is negligible but the method returns only 819 split-door instances, leading to low coverage over the full set of products. As α decreases, erroneous instances due to ϕ contribute to most of the magnitude of the error bars shown in Figure 7c. We observe that $\alpha = 0.95$ offers a good compromise: error bounds are within 1 percentage point and we obtain more than 7,000 split-door instances. The corresponding causal CTR estimate is 2.6%, with the error bars spanning $(2.0\%, 2.7\%)$. It is instructive to compare this to the naive observational estimate taken by computing the click-through rate

[6]Note that the error bars are asymmetrical; we expect erroneous split-door instances to drive the causal estimate up from its true value, under the assumption that demand for the two products are positively correlated with each other, as argued in Section 3.1.

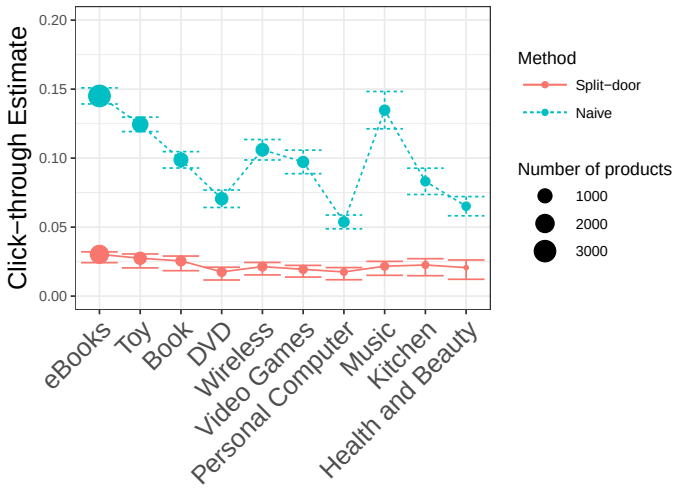
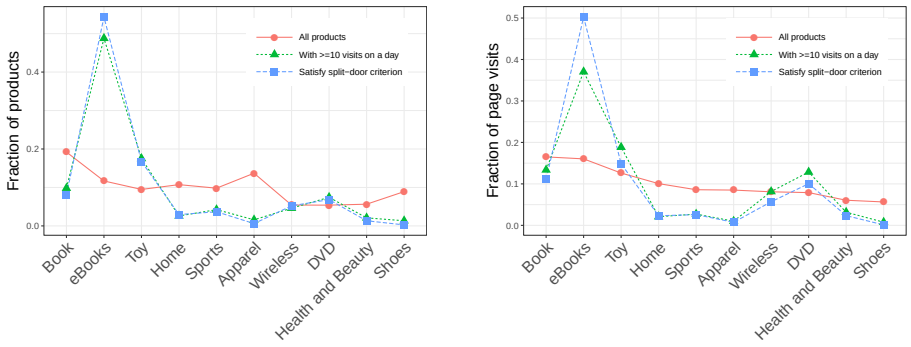


Fig 8: Comparison of causal CTR with naive observational CTR for products that satisfy the split-door criterion. Top 10 product categories are based on the number of products in the split-door sample.

across all focal and recommended product pairs, which produces an estimate of 9.6%. As in past work, this result shows that correlated demand results in an overestimate of the causal effect (Sharma, Hofman and Watts, 2015).

Furthermore, we can break these estimates down by the different product categories present on Amazon.com. Figure 8 shows the variation of $\hat{\rho}$ with the most popular categories, at a nominal threshold of 0.95. For the set of focal products that satisfy the split-door criterion, we also compute the naive observational estimate for CTR. We see substantial variation in the naive estimate, ranging from 14% on *e-Books* to 5% on *Personal Computer*. However, when we use the split-door criterion to compute estimates, we find that the causal CTR for all product categories lies below 5%. These results indicate that the naive observational estimate overstates the causal impact by anywhere from two- to five-fold across different product categories.

There are two clear advantages to the split-door criterion compared to past approaches for estimating the causal impact of recommender systems. First, we are able to study a larger fraction of products compared to instrumental variable approaches that depend on single-source variations (Carmi, Oestreicher-Singer and Sundararajan, 2012) or restricting our attention to mining only shocks in observational data (Sharma, Hofman and Watts, 2015). On the same dataset, the shock-based method in Sharma, Hofman and Watts (2015) identified valid instances on 4,000 unique focal products, while



(a) (b)

Fig 9: Distribution of products and total visits over product categories. Subset of focal products that satisfy the split-door criterion are near identical to the set of all products with at least 10 visits on any day, except for eBooks. Fraction of page visits to those focal products show more variation, but the overall distributions are similar.

the split-door criterion finds instances for over 5,000 unique focal products at $\alpha = 0.95$, and over 11,000 at $\alpha = 0.80$. Second, the split-door criterion provides a principled method to select valid instances for analysis by tuning α, the desired significance level, while also allowing for a estimate of the fraction of falsely accepted instances, ϕ.

That said, as with any observational analysis, the results rely on certain assumptions and so, of course, come with qualifications. Although the split-door criterion yields valid estimates of the causal impact of recommendations for the time periods where product pairs are found to be statistically independent, it is important to emphasize that the split-door time intervals are not selected at random, thus violating the *as-if-random* (Angrist, Imbens and Rubin, 1996) assumption powering generalizability for natural experiments. As a result, care must be taken to extrapolate these estimates to all products on Amazon.com. Fortunately, in this instance the split-door criterion covers a broad range of the observed data. For example, on Amazon.com, products with at least one valid split-door time-interval span multiple product categories and cover nearly a quarter of all focal products in the dataset at $\alpha = 0.95$. As an additional robustness check, we looked at the distribution over product categories for products identified by the split-door criterion and compared this to the same distribution for all products in the population. For comparison, we apply the same popularity filter—at least 10 page views on at least one day—to the dataset with all products.

Figure 9a shows that the distribution of products analyzed by the split-door criterion across different product categories is nearly identical to the overall set of products, except for eBooks which are over-represented in valid split-door instances. Figure 9b shows a similar result for the number of page visits across different product categories, with slightly higher deviation from the full sample of products. Although these results do not necessarily imply that the as-if-random assumption is satisfied (indeed it is very likely not satisfied) they do indicate that the split-door criterion at least allows us to estimate causal effects over a diverse sample of popular product categories, which is a clear improvement over past work (Carmi, Oestreicher-Singer and Sundararajan, 2012; Sharma, Hofman and Watts, 2015).

4. Discussion. In this paper we have presented a method for computing the causal effect of a variable X on another variable Y_R whenever we have an additional variable Y_D which follows some testable conditions, and have shown its application in estimating the causal impact of a recommender system. We now discuss some of the advantages and also limitations of the split-door criterion with respect to validity and generalizability of the estimate. We suggest guidelines to ensure proper use of the criterion and discuss other applications for which it might be used.

4.1. *Guidelines for using the criterion.* As with any non-experimental method for causal inference, the split-door criterion rests on certain assumptions, in particular: *independence* (of X and Y_D), *connectedness* (i.e. non-zero casual effect of U_Y on Y_D), and *unconfoundedness* (of X and Y_R). Here we give guidelines for reasoning about these assumptions and checking the robustness and sensitivity of estimates.

The Independence assumption is a standard assumption for observational causal inference. Barring coincidental equality of parameters such that the effect of unobserved confounders on X and Y_D cancel out, the independence assumption is likely to be satisfied. Nonetheless we encourage researchers to think carefully about this assumption in applying the criterion in other domains. Depending on the application it may be possible to rule out such cancellations. For example, in our recommendation system study we expect demand for the focal and recommended product to be correlated. Therefore, the causal effect of demand on both products is expected to be directionally identical, and hence cancellation becomes impossible.

The Connectedness assumption is potentially more restrictive, but can justified in domains where the measurements Y_R and Y_D are additive components of the same outcome Y. That said, it remains an untestable assumption where, once again, domain knowledge should be used to assess its

plausibility. For instance, even when Y_R and Y_D are additive components, in some isolated cases, U_{Yr} may not be connected to Y_D at all. In a recommender system this can happen when the population of customers who browse product j directly is completely independent from users who visit product i. In such a scenario, the split-door criterion would be invalid. We note, however, that this situation can arise only in the (unlikely) event that there is no relationship between demand for a product from users who visit the same product on the same website, just through different means. When there is even a small overlap between the demands of the two populations, the split-door criterion will again be valid and depending on the precision of the statistical independence condition, can also be applied empirically.

A key advantage of the split-door criterion is that once these two assumptions are met, it provides a statistical test for verifying its last and most critical assumption: unconfoundedness of X and Y_R. At the same time, statistical independence tests used in applying the split-door criterion often have their own free parameters. Any such parameters should be varied to check the sensitivity of estimates to these choices, as in Figures 7b and 7c.

Finally, after identifying and estimating the effect of interest, we need to consider how *useful* it is for practical applications. As remarked earlier and demonstrated in our recommender system application, the split-door criterion is capable of capturing the local average causal effect for a large sample of the dataset that satisfies the required independence assumption ($X \perp\!\!\!\perp Y_D$). The argument has been made that such local estimates are indeed useful in themselves (Imbens, 2010). That said, the sample may not be representative of the entire population, and so one must always be careful to qualify an extension of the split-door estimate to the general population. Naturally, the more instances discovered by the method, the more likely the estimate is to be of general use. Additionally, we recommend that researchers perform checks to compare the distribution of any available covariates to check for differences between the general population and instances that pass the split-door criterion.

4.2. *Potential applications of the split-door criterion.* The key requirement of the split-door criterion is that the outcome variable must comprise two distinct components: one that is potentially affected by the cause, and another that is not directly affected by it. In addition, we should have sufficient reason to believe that the two outcome components share common causes (i.e. the Connectedness assumption must be satisfied), and that one of outcome variables can be shown to be independent of the cause variable (i.e. the Independence assumption must be satisfied). These might seem like

overly restrictive assumptions that limit applicability of the criterion, but in this section we argue that there are in fact many interesting cases where the split-door criterion can be employed.

As we have already noted, recommendation systems such as Amazon's are particular well-suited to these conditions, in large part because Y_D has a natural interpretation of "direct traffic", or any traffic that is not caused by a particular recommendation. Likewise the criterion can be easily applied to other online systems that automatically log user visits, such as in estimating the effect of ads on search engines or websites. Somewhat more broadly, time series data in general may be amenable to the split-door criterion, in part because different components of the outcome occurring at the same time are more likely to be correlated than components that share other characteristics, and in part because time series naturally generate many observations on the input and output variables, which permits convenient testing for independence.

For example, consider the problem of estimating the effect of social media on news consumption. There has been recent interest (Flaxman, Goel and Rao, 2013) in how social media websites such as Facebook impact the news that people read, especially through algorithmic recommendations such as those for "Trending news". Given time series data for user activity on a social media website and article visits from news website logs, we can use the split-door criterion to estimate the effect of social media on news reading. Most websites record the source of each page visit, so obtaining two components for the outcome—visits to an article through through social media and through other means—should be straightforward. Here Y_R would correspond to the visits that are referred from social media, and Y_D would be all other direct visits to the news article. Whenever people's social media usage is not correlated with direct visits to a news article, we can identify the causal effect of social media on news consumption. Similar analysis can be applied to problems such as estimating the effect of online popularity of politicians on campaign financing or the effect of television advertisements on purchases.

Finally, although we have focused on online settings for which highly granular time series data is often collected by default, we note that there is nothing intrinsic to the split-door criterion that prevents it from being applied offline. Suppose, for example, that a brick and mortar store routinely sends discount coupons for one of its products to some of its customers. The split-door criterion could easily be used to estimate the causal effect of giving away discounts on product purchases: 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 (i.e. without a discount). U_Y represents common demand for the product that affects both Y_R and Y_D, as well as whether it was chosen as a discounted product (X). U_X represents all the other factors, independent of product demand U_Y, that may have gone into the selection of products to give away discounts for. More generally, the split-door criterion could be used to estimate the impact of the effect of targeted advertising on product sales, or in any context where demand for a given product can be differentiated into more than one channel.

5. Conclusion. In closing we note that the split-door criterion is just one example of a more general class of methods that adopt a data-driven approach to causal discovery (Jensen et al., 2008; Grosse-Wentrup et al., 2016; Sharma, Hofman and Watts, 2015; Cattaneo, Frandsen and Titiunik, 2015). As we have noted, data-driven methods have important advantages over traditional methods for exploiting natural variation—allowing inference to be performed on much larger and more representative samples—while also being far less susceptible to unobserved confounders than back-door identification strategies. As the volume and variety of fine-grained data continues to grow, we expect these methods to increase in popularity and to raise numerous questions regarding their theoretical foundations and practical applicability.

APPENDIX A: ESTIMATING THE FRACTION OF ERRONEOUS SPLIT-DOOR INSTANCES

Let the expected fraction of erroneous X-Y_D pairs—split-door instances—returned by the method be ϕ. In the terminology of multiple testing, ϕ refers to the *False Non-Discovery Rate (FNDR)* (Delongchamp et al., 2004). This is different from the more commonly used False Discovery Rate (FDR) (Farcomeni, 2008), since we deviate from standard hypothesis testing by looking for split-door instances that have a p-value higher than a pre-determined threshold. Given m hypothesis tests and a significance level of α, the false non-discovery rate ϕ for the split-door method can be characterized as

$$\phi_\alpha \leq \frac{(1-\alpha)\pi_{dep} m}{W_\alpha}, \tag{A.1}$$

where π_{dep} is the fraction of actually dependent X-Y_D instances in the dataset and W_α is the observed number of X-Y_D instances returned by the method at level α.

The above estimate can be derived using the framework proposed by Storey (2002) under two assumptions. The first is that the that the distribution of p-values under the null hypothesis is uniform, and the second is that the distribution of p-values under the alternative hypothesis is stochastic smaller than the uniform distribution. Let the number of invalid instances found using the split-door method be T. Then, by definition, the false non-discovery rate can be written as:

$$\phi_\alpha = \mathbf{E}\left[\frac{T}{W}\middle| W > 0\right].$$

Since the alternative distribution is stochastically smaller than uniform, we can arrive at an upper bound by replacing T by the expected number of split-door instances if the alternative distribution were uniform, $(1-\alpha) * m_{dependent} = (1-\alpha) * \pi_{dep} * m$, giving:

$$\phi_\alpha \leq \frac{(1-\alpha) * \pi_{dep} m}{W_\alpha}.$$

Here π_{dep} is unknown, so it needs to be estimated. A common approach is to estimate the fraction of actually independent instances or null hypotheses π_{indep} and then use $\pi_{dep} = 1 - \pi_{indep}$ (Delongchamp et al., 2004). For robustness, we suggest using multiple procedures to estimate π_{indep} and verify sensitivity of results to the choice of π_{indep}. In this paper, we use two different estimates, derived from Storey and Tibshirani (2003); Storey (2002) (*Storey's* estimate) and Nettleton et al. (2006); Liang and Nettleton (2012) (*Nettleton's* estimate).

Storey's estimate is defined as

$$\hat{\pi}_{indep} = \frac{W_\lambda}{m(1-\lambda)},$$

where $\lambda \in [0, 1)$ is a tunable parameter—similar in interpretation to α—and W_λ is the number of hypothesis tests having a p-value higher than λ. The choice of λ involves a bias-variance tradeoff, with $\lambda = 0.5$ being a common choice, as in the SAM software developed by Storey and Tibshirani (2003).

Nettleton's estimate, on the other hand, chooses the effective value of λ adaptively, based on the observed p-value distribution. First, the p-value distribution is summarized in a histogram containing B bins. Then, a threshold λ is chosen as the index (I) corresponding to the left-most bin whose count fails to exceed the average count of the bins to its right. This results in the following estimate, where $\lambda = (I-1)/B$:

$$\hat{\pi}_{indep} = \frac{W_\lambda}{m(1-\lambda)} = \frac{W_\lambda}{m(1-\frac{I-1}{B})}.$$

Applying each of these to the $m = 114,469$ focal and recommended product pairs analyzed in Section 3 allows us to estimate the true number of dependent X-Y_D pairs in the dataset, π_{dep}. At $\alpha = 0.95$, both methods give very similar results ($\pi_{dep,Storey} = 0.184$, $\pi_{dep,Nettleton} = 0.187$), and so we use $\pi_{dep} = 0.187$ in our analysis.

APPENDIX B: CHARACTERIZING ERROR IN THE SPLIT-DOOR ESTIMATE

The split-door causal estimate is defined as the mean of all estimates for each focal product in Section 3.3. Here we characterize the error in this estimate. The key idea is that the error comes from two components: the first due to some erroneously identified split-door instances, and the second due to natural variance in estimating the mean. For a significance level α of the independence test, let the number of split-door instances be W and the number of unique focal products be N. Then the mean estimate can be written as:

$$\hat{\rho} = \frac{\sum_{i\tau} \hat{\rho}_{i\tau}}{N}$$

For an expected number of $\phi W = \phi' N$ instances, the method may have erroneously concluded that the focal and recommended product are independent.[7] Therefore, the corresponding estimate for those focal products can be expanded as:

$$\hat{\rho}_{i\tau} = \rho_{i\tau}^{causal} + \eta_{i\tau}$$

where η refers to the click-through rate due to correlated demand between the focal and recommended product. Thus, the overall mean estimate can be written as:

$$\hat{\rho} = \frac{\sum_{i\tau \in A} \rho_{i\tau}^{causal} + \sum_{i\tau \in B} (\rho_{i\tau}^{causal} + \eta_{i\tau})}{N} = \frac{\sum_{i\tau} \rho_{i\tau}^{causal}}{N} + \frac{\sum_{i\tau \in B} \eta_{i\tau}}{N}$$

where A and B refer to sets of products with valid and erroneous split-door estimates respectively ($|A| = (1 - \phi')N, |B| = \phi' N$).

Comparing this to the true ρ^{causal},

$$\rho^{causal} - \hat{\rho} = (\rho^{causal} - \bar{\rho}^{causal}) - \frac{\sum_{i\tau \in B} \eta_{i\tau}}{N}$$

[7]Note that we generously apply ϕ to the number of split-door instances, assuming that an aggregated estimate for a focal product is invalid if split-door instance for *any* recommended product is invalid. In general, the expected number of invalid $\rho_{i\tau}$ estimates should be less than or equal to ϕW, since a focal product may have more than one recommended product that corresponds to an invalid split-door instance.

where the first term of the RHS corresponds to error due to sampling variance, and the second term corresponds to error due to ϕ. We estimate these terms below.

Error due to ϕ. Using the argument in Section 3.1, we can assume that the overall effect of U_Y on Y_R is positive (without stipulating it for each individual instance). This means that the term due to correlated demand is positive, $\sum_{i\tau \in B} \eta_{i\tau} \geq 0$. Further, the maximum value of $\eta_{i\tau}$ is attained when all the observed click-throughs are due to correlated demand ($\eta_{i\tau} = \hat{\rho}_{i\tau}$). Under this assumption,

$$0 \leq \frac{\sum_{i\tau \in B} \eta_{i\tau}}{N} \leq \frac{\rho_{maxsum}}{N}$$

where ρ_{maxsum} corresponds to the maximum sum of any subset of $\phi' N$ $\hat{\rho}_{i\tau}$ values. An approximate estimate can be derived using $\hat{\rho}$—the empirical mean over all N values of $\rho_{i\tau}$—leading to $\rho_{maxsum} \approx \phi' N \hat{\rho}$.

Error due to natural variance. We characterize this error by the 99% confidence interval for the mean estimate, given by $2.58 * \frac{\hat{\sigma}}{\sqrt{N}}$, where $\hat{\sigma}$ is the empirical standard deviation.

Combining these two, the resultant interval for the split-door estimate is:

$$(\hat{\rho} - \frac{\rho_{maxsum}}{N} - 2.58\frac{\hat{\sigma}}{\sqrt{N}}, \hat{\rho} + 2.58\frac{\hat{\sigma}}{\sqrt{N}}) \tag{B.1}$$

The above interval demonstrates the bias-variance tradeoff in choosing the nominal significance level for the independence test and the corresponding ϕ. At high nominal significance level α, bias due to ϕ is expected to be low but variance of the estimate may be high due to low N. Conversely, at low values of α, variance will be lower but ϕ is expected to be higher because we accept many more split-door instances.

SUPPLEMENTARY MATERIAL

Supplement A: Code for split-door criterion (http://www.github.com/amit-sharma/splitdoor-causal-criterion). We provide an R package for reproducing results and applying split-door criterion to new applications.

REFERENCES

AGRESTI, A. (1992). A survey of exact inference for contingency tables. *Statistical Science* 131–153.

AGRESTI, A. (2001). Exact inference for categorical data: recent advances and continuing controversies. *Statistics in medicine* **20** 2709–2722.

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association* **91** 444–455.

CARMI, E., OESTREICHER-SINGER, G. and SUNDARARAJAN, A. (2012). Is Oprah contagious? Identifying demand spillovers in online networks. *NET Institute Working Paper* **10-18**.

CATTANEO, M. D., FRANDSEN, B. R. and TITIUNIK, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference* **3** 1–24.

CATTANEO, M. D., TITIUNIK, R. and VAZQUEZ-BARE, G. (2017). Comparing inference approaches for RD designs: A reexamination of the effect of head start on child mortality. *Journal of Policy Analysis and Management*.

DE SIQUEIRA SANTOS, S., TAKAHASHI, D. Y., NAKATA, A. and FUJITA, A. (2014). A comparative study of statistical methods used to identify dependencies between gene expression signals. *Briefings in Bioinformatics* **15** 906-918.

DELONGCHAMP, R. R., BOWYER, J. F., CHEN, J. J. and KODELL, R. L. (2004). Multiple-Testing Strategy for Analyzing cDNA Array Data on Gene Expression. *Biometrics* **60** 774–782.

DUNNING, T. (2012). *Natural experiments in the social sciences: a design-based approach*. Cambridge University Press.

FARCOMENI, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical methods in medical research* **17** 347–388.

FISKE, S. T. and HAUSER, R. M. (2014). Protecting human research participants in the age of big data. *Proceedings of the National Academy of Sciences* **111** 13675-13676.

FLAXMAN, S., GOEL, S. and RAO, J. M. (2013). Ideological segregation and the effects of social media on news consumption. *Available at SSRN 2363701*.

GRAU, J. (2009). Personalized Product Recommendations: Predicting Shoppers' Needs.

GROSSE-WENTRUP, M., JANZING, D., SIEGEL, M. and SCHÖLKOPF, B. (2016). Identification of causal relations in neuroimaging data with latent confounders: An instrumental variable approach. *NeuroImage* **125** 825–833.

IMBENS, G. W. (2010). Better LATE than nothing. *Journal of Economic Literature* **48**.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

JENSEN, D. D., FAST, A. S., TAYLOR, B. J. and MAIER, M. E. (2008). Automatic identification of quasi-experimental designs for discovering causal knowledge. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* 372–380. ACM.

LEWIS, R. A., RAO, J. M. and REILEY, D. H. (2011). Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of the 20th international conference on World wide web* 157–166. ACM.

LIANG, K. and NETTLETON, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **74** 163-182.

LYDERSEN, S., PRADHAN, V., SENCHAUDHURI, P. and LAAKE, P. (2007). Choice of test for association in small sample unordered r× c tables. *Statistics in medicine* **26** 4328–4343.

MEALLI, F. and PACINI, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical As-*

sociation **108** 1120–1131.
Morgan, S. L. and Winship, C. (2014). *Counterfactuals and causal inference*. Cambridge University Press.
Mulpuru, S. (2006). *What you need to know about Third-Party Recommendation Engines*. Forrester Research.
Nettleton, D., Hwang, J. T. G., Caldo, R. A. and Wise, R. P. (2006). Estimating the number of true null hypotheses from a histogram of p values. *Journal of Agricultural, Biological, and Environmental Statistics* **11** 337.
Paninski, L. (2003). Estimation of entropy and mutual information. *Neural computation* **15** 1191–1253.
Pearl, J. (2009). *Causality*. Cambridge University Press.
Pethel, S. D. and Hahs, D. W. (2014). Exact test of independence using mutual information. *Entropy* **16** 2839–2849.
Phan, T. Q. and Airoldi, E. M. (2015). A natural experiment of social network formation and dynamics. *Proceedings of the National Academy of Sciences* **112** 6595–6600.
Ricci, F., Rokach, L. and Shapira, B. (2011). *Introduction to recommender systems handbook*. Springer.
Rosenzweig, M. R. and Wolpin, K. I. (2000). Natural" natural experiments" in economics. *Journal of Economic Literature* **38** 827–874.
Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press.
Sharma, A., Hofman, J. M. and Watts, D. J. (2015). Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation* 453–470.
Spirtes, P., Glymour, C. N. and Scheines, R. (2000). *Causation, prediction, and search*. MIT Press.
Steuer, R., Kurths, J., Daub, C. O., Weise, J. and Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* **18** S231–S240.
Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 479–498.
Storey, J. D. and Tibshirani, R. (2003). *SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays* In *The Analysis of Gene Expression Data: Methods and Software* 272–290. Springer New York, New York, NY.
Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* **25** 1.
Székely, G. J., Rizzo, M. L., Bakirov, N. K. et al. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics* **35** 2769–2794.

641 Ave. of the Americas
New York, NY USA 10011
E-mail: amshar@microsoft.com
jmh@microsoft.com
duncan@microsoft.com