From Optimizing Engagement to Measuring Value

Smitha Milli*
UC Berkeley
smilli@berkeley.edu

Luca Belli
Twitter
lbelli@twitter.com

Moritz Hardt†
UC Berkeley
hardt@berkeley.edu

Abstract

Most recommendation engines today are based on predicting user engagement, e.g. predicting whether a user will click on an item or not. However, there is potentially a large gap between engagement signals and a desired notion of *value* that is worth optimizing for. We use the framework of measurement theory to (a) confront the designer with a normative question about what the designer values, (b) provide a general latent variable model approach that can be used to operationalize the target construct and directly optimize for it, and (c) guide the designer in evaluating and revising their operationalization. We implement our approach on the Twitter platform on millions of users. In line with established approaches to assessing the *validity* of measurements, we perform a qualitative evaluation of how well our model captures a desired notion of "value".

1 Introduction

Most recommendation engines today are based on predicting user engagement, e.g. predicting whether a user will click an item or not. However, there is potentially a large gap between engagement signals and a desired notion of *value* that is worth optimizing for [Ekstrand and Willemsen, 2016]. Just because a user engages with an item doesn't mean they value it. A user might reply to an item because they are angry about it, or click an item in order to gain more information about it [Wen et al., 2019], or watch addictive videos out of temptation.

It is clear that engagements provide some signal for "value", but are not equivalent to it. Further, different types of engagement may provide differing levels of evidence for value. For example, if a user explicitly likes an item, we are more likely to believe that they value it, compared to if they had merely clicked on it. Ideally, we want the objective for our recommender system to take engagement signals into account, but only insofar as they relate to a desired notion of "value". However, directly specifying such an objective is a non-trivial problem. Exactly how much should we rely on likes versus clicks versus shares and so on? How do we evaluate whether our designed objective captures our intended notion of "value"?

1.1 Our contributions

We make three primary contributions.

1. We propose measurement theory as a principled approach to aggregating engagement signals into an objective function that captures a desired notion of "value". The resulting objective function can be optimized from data, serving as a plug-in replacement for the ad-hoc objectives typically used in engagement optimization frameworks.

2. Our approach is based on the creation of a latent variable model that relates value to various observed engagement signals. We devise a new identification strategy for the latent variable model tailored to the

[*] Work done while the author was an intern at Twitter.

[†] MH is a paid consultant at Twitter. Work performed while consulting for Twitter.

intended use case of online recommendation systems. Our identification strategy needs only a single robust engagement signal for which we know the conditional probability of value given the signal.

3. We implemented our approach on the Twitter platform on millions of users. In line with an established validity framework for measurement theory, we conduct a qualitative analysis of how well our model captures "value".

1.2 Measurement theory and latent variable models

The framework of *measurement theory* [Hand, 2004, Jackman, 2009] is widely used in the social sciences as a guide to measuring *unobservable theoretical constructs* like "quality of life", "political ideology", or "socio-economic status". Under the measurement approach, theoretical constructs are operationalized as latent variables, which are related to observable data through a latent variable model (LVM). '

Similarly, we treat the "value" of a recommendation as a theoretical construct, which we operationalize as a (binary) latent variable V. We represent the LVM as a a *Bayesian network* [Pearl, 2009] that contains V as well as each of the possible types of user engagements (clicks, shares, etc). The structure of the Bayesian network allows us to specify conditional independences between variables, enabling us to capture dependencies like e.g. needing to click an item before replying to it.

Under the measurement approach, the ideal objective becomes clear: $\mathbb{P}(V = 1 \mid \texttt{Behaviors})$ - the probability the user values the item given their engagements with it. Such an objective uses all engagement signals, but only insofar provide evidence of Value V. If we can identify $\mathbb{P}(V = 1 \mid \texttt{Behaviors})$, then it can be used as a drop-in replacement for any objective that scores items based on engagement signals.

Our key insight is that we can identify $\mathbb{P}(V \mid \texttt{Behaviors})$ — the probability of Value given *all* behaviors — through the use of a single *anchor variable* A for which we know $\mathbb{P}(V = 1 \mid A = 1)$. The anchor variable, together with the structure of the Bayesian network, is what gives "value" its meaning. Through the choice of the anchor variable and the structure of the Bayesian network, the designer has the flexibility to give "value" subtly different meanings.

Recommendation engines have natural candidates for anchor variables: strong, explicit feedback from the user. For example, strong negative feedback could include downvoting or reporting a content item, or blocking another user. Strong positive feedback could be explicitly liking or upvoting an item. For negative feedback, we make the assumption that $\mathbb{P}(V = 1 \mid A = 1) = \epsilon$ for $\epsilon \approx 0$, while for positive feedback we make the assumption that $\mathbb{P}(V = 1 \mid A = 1) = 1 - \epsilon$.

1.3 A case study on the Twitter platform

We implemented our approach on the Twitter platform on millions of users. On Twitter, there are numerous user behaviors: clicks, favorites, retweets, replies, and many more. It would be difficult to directly specify an objective that properly trades-off all these behaviors. Instead, we identify a natural anchor variable. On Twitter, users can give explicit feedback on tweets by clicking "See less often" (SLO) on them. We use SLO as our anchor and assume that the user does not value tweets they click "See less often" on. After specifying the anchor variable and the Bayesian network, we are able to learn $\mathbb{P}(V \mid \texttt{Behaviors})$ from data.

The model automatically learns a natural ordering of which behaviors should provide stronger evidence for Value V, e.g. $\mathbb{P}(V = 1 \mid \texttt{Retweet} = 1) > \mathbb{P}(V = 1 \mid \texttt{Reply} = 1) > \mathbb{P}(V = 1 \mid \texttt{Click} = 1)$. Furthermore, it learns complex inferences about the evidence provided by *combinations* of behavior. Such inferences would not be possible under the standard approach, which uses a linear combination of behaviors as the objective.

Unlike other work on recommender systems, we do not evaluate through engagement metrics. If we believe that engagement is not the same as the construct "value", then we cannot evaluate our approach merely by reporting engagement numbers. Instead, we must take a more holisitc approach. We discuss established approaches to assessing the *validity* [AERA, APA, NCME, 2014, Messick, 1987, Reeves and Marbach-Ad, 2016] of a measurement, and explain how they translate to the recommender system setting by using Twitter as an example.

2 Related work

In the social sciences, especially in psychology, education, and political science, measurement theory [Hand, 2004] has long been used to operationalize constructs like "personality", "intelligence", "political ideology", etc. Often the operationalization of such constructs is heavily contested, and many types of evidence for validity and reliability are used to evaluate the match between a construct and its operationalization [Messick, 1987, AERA, APA, NCME, 2014].

Recently, Jacobs and Wallach [2019] introduced the language of measurement in the context of computer science. They argue that many harms effected by computational systems are the direct result of a mis-match between a theoretical construct and its operationalization. In the context of recommender systems, many have argued that the engagement metrics used in practice are a poor operationalization of "value" [Ekstrand and Willemsen, 2016].

We use measurement theory as a principled way to disentangle latent value from observed engagement. We provide a general latent variable model approach in which an *anchor variable* provides the key link between the latent variable and the observed behaviors. The term anchor variable has been used been used in various ways in prior work on factor models [Arora et al., 2012, 2013, Halpern et al., 2016a,b]; our usage is most similar to [Halpern et al., 2016b]. Our use of the anchor variable is also similar to the use of a *proxy variable* to identify causal effects under unobserved confounding [Pearl, 2010, Kuroki and Pearl, 2014].

3 Identification of the LVM with anchor A

We now describe our general approach to operationalizing a target construct through a latent variable model (LVM) with an *anchor variable*. We operationalize the construct for value through a LVM in which the construct is represented through an unobserved, binary latent variable V that the other binary, observed behaviors provide evidence for. We assume there is one observed behavior, an *anchor variable* A, which we know $\mathbb{P}(V = 1 \mid A = 1)$ for. We represent all other observed behaviors in the binary random vector $\mathbf{B} = (B_1, \ldots, B_n)$. We refer to A as an anchor variable because it will provide the crucial link to identifying $\mathbb{P}(V \mid A, \mathbf{B})$. In other words, it will *anchor* the other observed behaviors $\mathbf{B}$ to Value V.

We represent the LVM as a Bayesian network. A Bayesian network is a directed acyclic graph (DAG) that graphically encodes a factorization of the joint distribution of the variables in the network. In particular, the DAG encodes all conditional independences among the nodes through the d-separation rule [Pearl, 2009]. This is important because in most real-world settings, the observed behaviors have complex dependencies among each other (e.g. one may need to click on an item before replying to it). Through our choice of the DAG we can model both the dependencies among the observed behaviors as well as the dependence of the unobserved variable V on the observed behaviors.

Our goal is to determine $\mathbb{P}(V \mid A, \mathbf{B})$ so that it can later be used downstream as a target for optimization. We now discuss sufficient conditions for identifying the conditional distribution $\mathbb{P}(V \mid A, \mathbf{B})$. There are three assumptions on the anchor variable A that we will consider in turn.

Notation. We use $\text{Pa}(X)$ to denote the parents of a node X and use $\text{Pa}_{-V}(X) = \text{Pa}(X) \setminus V$ to denote all parents of X except for V.

Assumption 1 (Value-sensitive)**.** *For every realization b of the random vector $\mathbf{B}$, we have that $\mathbb{P}(A = 1 \mid \mathbf{B} = b, V = 1) \neq \mathbb{P}(A = 1 \mid \mathbf{B} = b, V = 0)$.*

Assumption 1 simply means that the anchor A carries signal about Value V, regardless of what the other variables $\mathbf{B}$ are.[1]

Assumption 2 (No children)**.** *The anchor variable A has no children.*

Since the anchor A is chosen to be a strong type of explicit feedback, it is usually the last type of behavior the user engages in on a content item (e.g. a "report" button that removes the content from the user's timeline), and thus, it typically makes sense to model A as having no children.

Assumption 3 (One-sided conditional independence)**.** *Let $\text{Pa}_{-V}(A)$ be all parents of A excluding V. Value V is independent from $\text{Pa}_{-V}(A)$ given that $A = 1$:*

$$\mathbb{P}(V = 1 \mid A = 1, \text{Pa}_{-V}(A)) = \mathbb{P}(V = 1 \mid A = 1)\,.$$

Assumption 3 means that when the user has opted to give feedback ($A = 1$), the level of information that feedback contains about Value V does not depend on the other parents of A. The assumption rests on the fact that A is a strong type of feedback that the user only provides when they are confident of their assessment.

3.1 Conditions for identification

The next theorem establishes that under A1, the distribution of observable behaviors $\mathbb{P}(A, \mathbf{B})$ and the conditional distribution $\mathbb{P}(A \mid V, \mathbf{B})$ are sufficient for identifying the conditional distribution, $\mathbb{P}(V \mid A, \mathbf{B})$. The proof uses a *matrix adjustment method* (Rothman et al., 2008; pg. 360) and is very similar to that in Pearl [2010], Kuroki and Pearl [2014].

Theorem 1. *Let V and A be binary random variables and let $\mathbf{B} = (B_1, \ldots, B_n)$ be a binary random vector. If A1 holds, then the distributions $\mathbb{P}(A, \mathbf{B})$ and $\mathbb{P}(A \mid V, \mathbf{B})$ uniquely identify the conditional distribution $\mathbb{P}(V \mid A, \mathbf{B})$.*

Proof. Since the conditional distribution $\mathbb{P}(V \mid A, \mathbf{B})$ is equal to $\frac{\mathbb{P}(\mathbf{B},V)\cdot\mathbb{P}(A|\mathbf{B},V)}{\mathbb{P}(A,\mathbf{B})}$, we can reduce the problem to determining the distribution $\mathbb{P}(\mathbf{B}, V)$. We can relate $\mathbb{P}(\mathbf{B}, V)$ to the given distributions, $\mathbb{P}(A, \mathbf{B})$ and $\mathbb{P}(A \mid \mathbf{B}, V)$, via the law of total probability:

$$\mathbb{P}(A, \mathbf{B}) = \sum_{v \in \{0,1\}} \mathbb{P}(\mathbf{B}, V = v)\mathbb{P}(A \mid \mathbf{B}, V = v)\,. \tag{1}$$

For every realization b of the random vector $\mathbf{B}$, we can write Equation 1 as $z^b = \mathbf{P}^b \mu^b$ where the matrix $\mathbf{P}^b \in [0, 1]^{2\times 2}$ and the vectors $\mu^b, z^b \in [0, 1]^2$ are defined as

$$\begin{aligned}
&\mathbf{P}^b_{i,j} = \mathbb{P}(A = i \mid \mathbf{B} = b, V = j) \quad \text{for} \quad i, j \in \{0, 1\}\,, \\
&\mu^b = [\mathbb{P}(\mathbf{B} = b, V = 0), \mathbb{P}(\mathbf{B} = b, V = 1)]^T\,, \\
&z^b = [\mathbb{P}(\mathbf{B} = b, A = 0), \mathbb{P}(\mathbf{B} = b, A = 1)]^T\,.
\end{aligned}$$

Determining the distribution $\mathbb{P}(\mathbf{B}, V)$ is equivalent to determining μ^b for all b. By Assumption 1, for all b we have $\mathbb{P}(A = 1 \mid B = b, V = 1) \neq \mathbb{P}(A = 1 \mid B = b, V = 0)$, which implies that the determinant of the matrix $\mathbf{P}^b$ is non-zero. Therefore, for all b, the vector μ^b is equal to $\mu^b = (\mathbf{P}^b)^{-1}z^b$. Thus, $\mathbb{P}(\mathbf{B}, V)$, and therefore the conditional distribution $\mathbb{P}(V \mid A, \mathbf{B})$, is identified by the given distributions. □

If we add Assumption 2, i.e. the anchor A has no children, then the distributions $\mathbb{P}(A, \mathbf{B})$ and $\mathbb{P}(A \mid \text{Pa}(A))$ are sufficient to identify $\mathbb{P}(V \mid A, \mathbf{B})$.

[1] When combined with Assumption 2, Assumption 1 simplifies to the condition $\mathbb{P}(A = 1 \mid \text{Pa}_{-V}(A) = z, V = 1) \neq \mathbb{P}(A = 1 \mid \text{Pa}_{-V}(A) = z, V = 0)$ for every realization z of $\text{Pa}_{-V}(A)$, the parents of A excluding V.

Corollary 1. *If the joint distribution $\mathbb{P}(V, A, \mathbf{B})$ is Markov*[2] *with respect to a DAG G in which A1 and A2 hold, then the distributions $\mathbb{P}(A, \mathbf{B})$ and $\mathbb{P}(A \mid \mathrm{Pa}(A))$ uniquely identify the conditional distribution $\mathbb{P}(V \mid A, \mathbf{B})$.*

Proof. In a Bayesian network, the *Markov blanket* for a variable X is the set of variables $\mathrm{MB}(X) \subseteq \mathcal{Z}$ that shield X from all other variables $\mathcal{Z}$ in the DAG, i.e. $\mathbb{P}(X \mid \mathcal{Z}) = \mathbb{P}(X \mid \mathrm{MB}(X))$ [Pearl, 2009]. The Markov blanket for a variable X consists of its parents, children, and parents of its children. Since the anchor A has no children, $\mathbb{P}(A \mid V, \mathbf{B}) = \mathbb{P}(A \mid \mathrm{MB}(A)) = \mathbb{P}(A \mid \mathrm{Pa}(A))$. Thus, by Theorem 1, $\mathbb{P}(A \mid \mathrm{Pa}(A))$, and $\mathbb{P}(A, \mathbf{B})$ identify the conditional distribution $\mathbb{P}(V \mid A, \mathbf{B})$ □

Finally, when we add Assumption 3, one-sided conditional independence, then the distributions $\mathbb{P}(V)$, $\mathbb{P}(A, \mathbf{B})$, $\mathbb{P}(V = 1 \mid A = 1)$, and $\mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V)$ are sufficient. The proof follows from Corollary 1 because, under Assumption 3, the distributions $\mathbb{P}(V = 1 \mid A = 1)$, $\mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V)$, and $\mathbb{P}(V)$ identify $\mathbb{P}(A \mid \mathrm{Pa}(A))$.

Corollary 2. *If the joint distribution $\mathbb{P}(V, A, \mathbf{B})$ is Markov with respect to a DAG G in which A1-3 hold, then $\mathbb{P}(V)$, $\mathbb{P}(A, \mathbf{B})$, $\mathbb{P}(V = 1 \mid A = 1)$, and $\mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V)$ uniquely identify the conditional distribution $\mathbb{P}(V \mid A, \mathbf{B})$.*

Proof. We will show that, under Assumption 3, the distributions $\mathbb{P}(V = 1 \mid A = 1)$, $\mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V)$, and $\mathbb{P}(V)$ identify $\mathbb{P}(A \mid \mathrm{Pa}(A))$. The proof then follows from Corollary 1.

We show that we can identify $\mathbb{P}(A \mid \mathrm{Pa}(A))$ by solving a set of linear equations. For short-hand let $p_{w,a,v} = \mathbb{P}(\mathrm{Pa}_{-V}(A) = w, A = a, V = v)$. For any realization w, by marginalizing over A and V, we can derive the following four equations for the four unknown probabilities $p_{w,0,0}$, $p_{w,0,1}$, $p_{w,1,0}$, $p_{w,1,1}$:

$$\mathbb{P}(\mathrm{Pa}_{-V}(A) = w, A = 0) = p_{w,0,0} + p_{w,0,1} \tag{2}$$
$$\mathbb{P}(\mathrm{Pa}_{-V}(A) = w, A = 1) = p_{w,1,0} + p_{w,1,1} \tag{3}$$
$$\mathbb{P}(\mathrm{Pa}_{-V}(A) = w, V = 0) = p_{w,0,0} + p_{w,1,0} \tag{4}$$
$$\mathbb{P}(\mathrm{Pa}_{-V}(A) = w, V = 1) = p_{w,0,1} + p_{w,1,1} \tag{5}$$

Note that the LHS of Equations 2 and 3 are given by $\mathbb{P}(A, \mathbf{B})$ and the LHS of Equations 4 and 5 are given by the prior $\mathbb{P}(V)$ and $\mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V)$.

From Assumption 3, one-sided conditional independence, we know that $\mathbb{P}(V = 1 \mid A = 1, \mathrm{Pa}_{-V}(A)) = \mathbb{P}(V = 1 \mid A = 1)$. Under one-sided conditional independence, the probability $p_{w,1,1}$ is determined by the given distributions:

$$\begin{aligned} p_{w,1,1} = \ & \mathbb{P}(A = 1) \cdot \mathbb{P}(\mathrm{Pa}_{-V}(A) = w \mid A = 1) \\ & \cdot \mathbb{P}(V = 1 \mid A = 1, \mathrm{Pa}_{-V}(A) = w) \\ = \ & \mathbb{P}(A = 1) \cdot \mathbb{P}(\mathrm{Pa}_{-V}(A) = w \mid A = 1) \\ & \cdot \mathbb{P}(V = 1 \mid A = 1) \,. \end{aligned} \tag{6}$$

Since $p_{w,1,1}$ is determined by the given distributions, so are $p_{w,0,0}$, $p_{w,1,0}$, and $p_{w,0,1}$, which can be solved for through Equations 2-5. Since this holds for any realization w, the distribution $\mathbb{P}(A, V, \mathrm{Pa}_{-V}(A)) = \mathbb{P}(A, \mathrm{Pa}(A))$ is determined, which by Collorary 1 means that the conditional distribution $\mathbb{P}(V \mid A, \mathbf{B})$ is determined. □

[2] A distribution $\mathbb{P}(X_1, \ldots, X_n)$ is said to be Markov with respect to a DAG G if it factorizes according to G, i.e. $\mathbb{P}(X_1, \ldots, X_n) = \prod_{i \in [n]} \mathbb{P}(X_i \mid \mathrm{Pa}(X_i))$.

3.2 Specifying the distributions for identification

Corollary 2 establishes that, under Assumptions 1-3, the distributions $\mathbb{P}(V)$, $\mathbb{P}(A, \mathbf{B})$, $\mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V)$, and $\mathbb{P}(A = 1 \mid V = 1)$ are sufficient to determine the conditional distribution $\mathbb{P}(V \mid A, \mathbf{B})$ for the LVM. Where do we get these distributions?

1. The distribution of observable nodes $\mathbb{P}(A, \mathbf{B})$ is estimated by the empirical distribution of observed data.
2. The distribution $\mathbb{P}(V)$ over the latent variable for value V is a prior distribution that is specified by the modeler. Recall that our goal with the LVM is to use $\mathbb{P}(V = 1 \mid \mathbf{B}, A)$ as an objective to optimize. Since the prior $\mathbb{P}(V)$ only has a scaling effect on $\mathbb{P}(V = 1 \mid \mathbf{B}, A)$, it does not matter greatly. We set $\mathbb{P}(V)$ to be uniform, i.e. $\mathbb{P}(V = 1) = 0.5$.
3. The conditional probability $\mathbb{P}(V = 1 \mid A = 1)$ is specified by our assumption on the the anchor variable. The probability $\mathbb{P}(V = 1 \mid A = 1)$ is set to ϵ where $\epsilon \approx 0$ if A is explicit negative feedback or to $1 - \epsilon$ if A is explicit positive feedback.
4. That leaves the distribution $\mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V)$. We estimate $\mathbb{P}(\mathrm{Pa}_{-V}(A) \mid A)$ heuristically using two sources of historical data that vary in their distribution of Value V. Suppose we have access to a dataset of historical recommendations $\mathcal{D}_R$ that were sent to users at random, as well as a dataset of historical recommendations that were algorithmically chosen, $\mathcal{D}_C$. The randomized and algorithmic datasets will have different distributions of valuable content, $\mathbb{P}_R(V)$ and $\mathbb{P}_C(V)$, and different distributions of observed behavior, $\mathbb{P}_R(A, \mathbf{B})$ and $\mathbb{P}_C(A, \mathbf{B})$. However, we assume that $\mathbb{P}(A, \mathbf{B} \mid V)$, the probability of the observed behavior given Value V, is the same between the two datasets.[3] The following equations then hold:

$$\begin{aligned}\mathbb{P}_R(\mathrm{Pa}_{-V}(A)) = \;& \mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V = 1)\mathbb{P}_R(V = 1) \\ & + \mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V = 0)\mathbb{P}_R(V = 0)\,, \qquad (7) \\ \mathbb{P}_C(\mathrm{Pa}_{-V}(A)) = \;& \mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V = 1)\mathbb{P}_C(V = 1) \\ & + \mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V = 0)\mathbb{P}_C(V = 0)\,. \qquad (8)\end{aligned}$$

   We specify $\mathbb{P}_R(V)$ and $\mathbb{P}_C(V)$ in an application-dependent way, but, generally, we assume the randomized dataset is lower value than the algorithmic one: $\mathbb{P}_R(V) < \mathbb{P}_C(V)$. Once we specify $\mathbb{P}_R(V)$ and $\mathbb{P}_C(V)$ and estimate $\mathbb{P}_R(A, \mathbf{B})$ and $\mathbb{P}_C(A, \mathbf{B})$ empirically, then we can solve Equations 7 and 8 to estimate $\mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V = 1)$. This is a heuristic approach that is appropriate for getting a rough estimate, but needs to be used with care. In practice, not all the differences between the randomized and algorithmic dataset can be explained by an intervention on Value V. For example, if the recommendation algorithm has historically been optimized for user clicks, then users in the algorithmic dataset may click on items more, but for reasons other than increased value.

3.3 Algorithm for identification

We now give more details on how we calculate the joint distribution $\mathbb{P}(V, A, \mathbf{B})$ given the distributions $\mathbb{P}(V)$, $\mathbb{P}(A, \mathbf{B})$, $\mathbb{P}(V = 1 \mid A = 1)$ and $\mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V)$. We use the structure of the Bayesian network to efficiently identify the joint distribution $\mathbb{P}(V, A, \mathbf{B})$ by fitting each factor $\mathbb{P}(X \mid \mathrm{Pa}(X))$ for every variable X.

1. The factor for V is given by the prior $\mathbb{P}(V)$.[4]
2. The factor for A, i.e. $\mathbb{P}(A \mid \mathrm{Pa}(A))$, can be identified from $\mathbb{P}(V)$, $\mathbb{P}(V = 1 \mid A = 1)$, and $\mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V)$ by solving a set of linear equations as in the proof of Corollary 2.

[3] If our DAG has Value V as a root node and can be interpreted as a causal Bayesian network [Pearl, 2009], then this is equivalent to assuming that the difference between the datasets corresponds to an intervention on V.

[4] Assuming that V is a root node, which is the case in any network we are interested in.

3. The factor for any behavior that does not have V as a parent is directly identified by the distribution of observable behaviors $\mathbb{P}(A, \mathbf{B})$.
4. The factors for the remaining behaviors which have V as a parent are fit through a *matrix adjustment method* (Rothman et al., 2008; pg. 360). In particular, note that

$$\begin{aligned}&\mathbb{P}(X = 1, \mathrm{Pa}_{-V}(X) = z, \mathrm{Pa}_{-V}(A) = w, A = a) = \\ &(\sum_{v \in \{0,1\}} \mathbb{P}(A = a \mid \mathrm{Pa}_{-V}(A) = w, V = v) \\ &\cdot \mathbb{P}(X = 1, \mathrm{Pa}_{-V}(X) = z, \mathrm{Pa}_{-V}(A) = w, V = v))\end{aligned}$$

We can also write the above equation in matrix form. Let $z_1, \ldots, z_m$ be all realizations of $\mathrm{Pa}_{-V}(X)$, and define the matrices $Q^w \in [0,1]^{2\times m}$, $R^w \in [0,1]^{2\times 2}$, $S^w \in [0,1]^{2\times m}$ as[5]

$$Q^w_{a,i} = \mathbb{P}(X = 1, \mathrm{Pa}_{-V}(X) = z_i, \mathrm{Pa}_{-V}(A) = w, A = a), \tag{9}$$

$$R^w_{av} = \mathbb{P}(A = a \mid \mathrm{Pa}_{-V}(A) = w, V = v), \tag{10}$$

$$S^w_{v,i} = \mathbb{P}(X = 1, \mathrm{Pa}_{-V}(X) = z_i, \mathrm{Pa}_{-V}(A) = w, V = v). \tag{11}$$

Then, $Q^w = R^w S^w$ and $S^w = (R^w)^{-1}Q^w$.[6] Let S be the marginalization over w: $\sum_w S^w = (R^w)^{-1}Q^w$. Then $S_{v,i} = \mathbb{P}(X = 1, \mathrm{Pa}_{-V}(X) = z_i, V = v)$. Thus, the factor for X is equal to $\mathbb{P}(X \mid \mathrm{Pa}_{-V}(X) = z_i, V = v) = S_{v,i}/\mathbb{P}(\mathrm{Pa}_{-V}(X) = z_i, V = v)$. We fit nodes with V as a parent in topological order, so that we can always calculate the denominator from previously fit factors.

4 Application to Twitter

We implemented our approach on the Twitter platform on millions of users. On Twitter, there are many kinds of user behaviors: clicks, replies, favorites, retweets, etc. The typical approach to recommendations would involve optimizing an objective that trades-off these behaviors, usually with linear weights. However, designing an objective is a non-trivial problem. How exactly should we weigh favorites compared to clicks or replies or retweets or any of the numerous other behaviors? It is difficult to assess whether the weights we chose match the notion of "value" we intended.

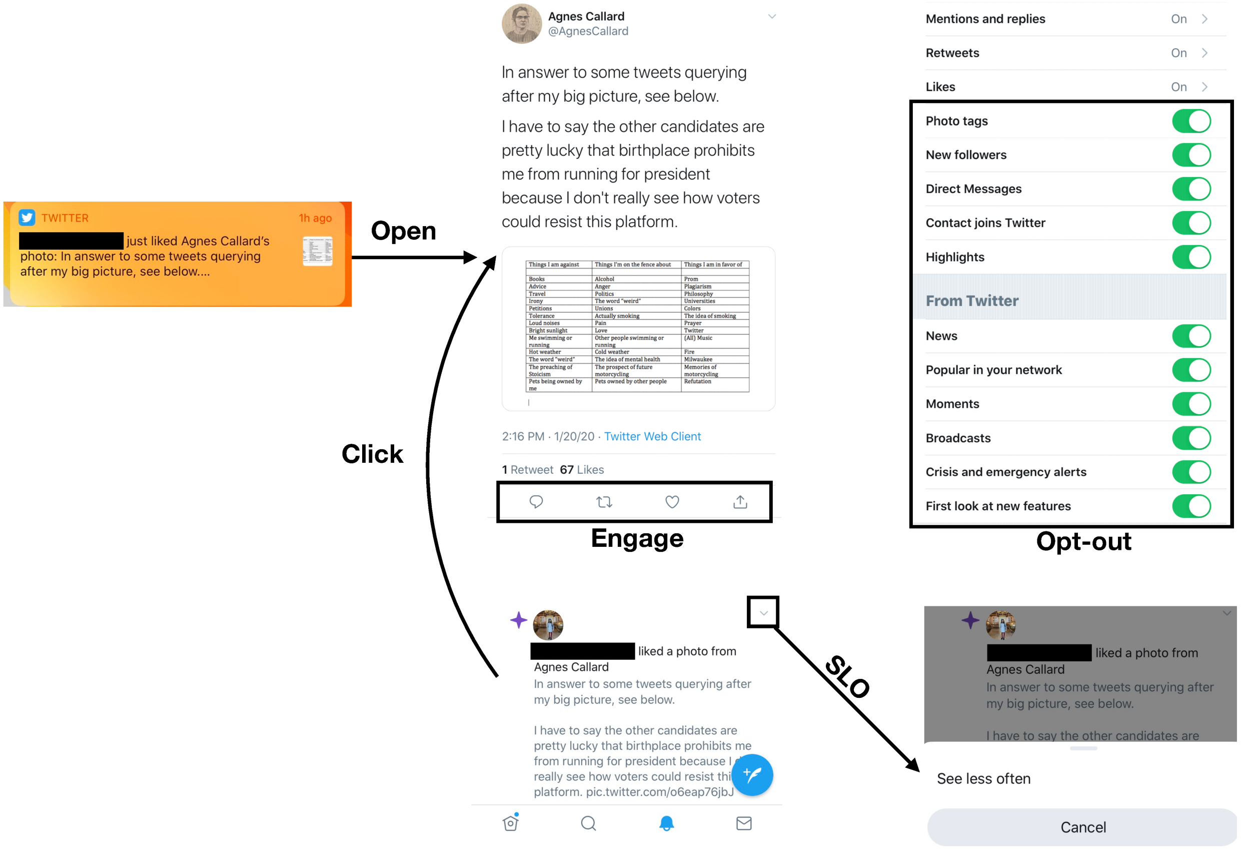
Furthermore, even supposing that we could manually specify the "correct" weights through laborious trial-and-error, the correct weights change over time. For example, after videos shared on Twitter began to auto-play, the signal of whether or not a user watched a video presumably became less relevant. The reality is that the objective is never static - how users interact with the platform is constantly changing, and the objective must change accordingly.

Our approach provides a principled solution to objective specification. We directly operationalize our intended construct "value" as a latent variable V. The meaning of Value V is defined by the Bayesian network and the *anchor variable* A, a behavior that we believe provides strong evidence for value or the lack of it. On Twitter, the user can provide strong, explicit feedback by clicking "See less often" (SLO) on a tweet. We use SLO as our anchor A and assume that if a user clicks "See less often" on a tweet, they do not value it: $\mathbb{P}(V = 1 \mid \texttt{SLO} = 1) = 0$.

Under this approach, there is no need to manually specify how all the behaviors should factor into the objective. Having operationalized Value, the ideal objective to use is clear: $\mathbb{P}(V = 1 \mid \mathbf{B}, A)$ - the probability of Value V given the observed behaviors. As discussed in Section 4, we can directly estimate $\mathbb{P}(V = 1 \mid \mathbf{B}, A)$

[5] If $\mathrm{Pa}_{-V}(X) \cap \mathrm{Pa}_{-V}(A) \neq \emptyset$ and $\mathrm{Pa}_{-V}(X) = z_i$ and $\mathrm{Pa}_{-V}(A) = w$ conflict, then simply set $Q^w_{0,i} = Q^w_{1,i} = 0$.

[6] R^w is invertible because of Assumption 1.



TWITTER 1h ago

[redacted] just liked Agnes Callard's photo: In answer to some tweets querying after my big picture, see below....

Agnes Callard
@AgnesCallard

In answer to some tweets querying after my big picture, see below.

I have to say the other candidates are pretty lucky that birthplace prohibits me from running for president because I don't really see how voters could resist this platform.

Things I am against	Things I'm on the fence about	Things I am in favor of
Books	Alcohol	Prom
Advice	Anger	Plagiarism
Travel	Politics	Philosophy
Irony	The word "weird"	Universities
Petitions	Unions	Colors
Tolerance	Actually smoking	The idea of smoking
Loud noises	Pain	Prayer
Bright sunlight	Love	Twitter
Me swimming or running	Other people swimming or running	(All) Music
Hot weather	Cold weather	Fire
The word "weird"	The idea of mental health	Milwaukee
The preaching of Stoicism	The prospect of future motorcycling	Memories of motorcycling
Pets being owned by me	Pets owned by other people	Refutation

2:16 PM · 1/20/20 · Twitter Web Client

1 Retweet 67 Likes

Mentions and replies On
Retweets On
Likes On
Photo tags
New followers
Direct Messages
Contact joins Twitter
Highlights

From Twitter

News
Popular in your network
Moments
Broadcasts
Crisis and emergency alerts
First look at new features

[redacted] liked a photo from Agnes Callard
In answer to some tweets querying after my big picture, see below.

I have to say the other candidates are pretty lucky that birthplace prohibits me from running for president because I d really see how voters could resist thi platform. pic.twitter.com/o6eap76jbJ

[redacted] liked a photo from Agnes Callard
In answer to some tweets querying after my big picture, see below.

I have to say the other candidates are

See less often

Cancel

Figure 1: A workflow of how users can interact with ML-based notifications on Twitter. To view the tweet, the user can either "open" the notification from the home screen on their phone or "click" on it from the notifications tab within the app. If the user sees the tweet from their notifications tab, they can also click "See Less Often" on it. Once the user has opened or clicked on the notification, they can engage with the tweet in many ways, e.g. replying, retweeting, or favoriting. At any point, the user can opt-out of notifications all-together.

from data. Furthermore, presuming that the anchor and structure of Bayesian network remain stable, we can regularly re-estimate the model with new data at any point, allowing us to account for change in user behavior on the platform.

The Bayesian network. We applied our approach to ML-driven notifications on Twitter. These notifications have various forms, e.g. "Users A, B, C just liked User Z's tweet", "User A just tweeted after a long time", or "Users A, B, C followed User Z". Figure 1 shows an example notification and how a user can interact with it. The Bayesian network in Figure 2 succinctly encodes the dependencies between different types of interactions users can have with notifications.[7]

Notifications are sent both to the user's home screen on their mobile phone, as well as to the notifications tab within the Twitter app. The user can start their interaction either by seeing the notification in their notification tab (`NTabView`), and then clicking on it (`Click`), or by seeing it as a the notification on their phone home screen and opening it from there directly (`Open`). After clicking or opening the notification, the

[7] The network can be interpreted as a *causal* Bayesian network [Pearl, 2009], although for our purposes, we do not strictly need the causal interpretation.

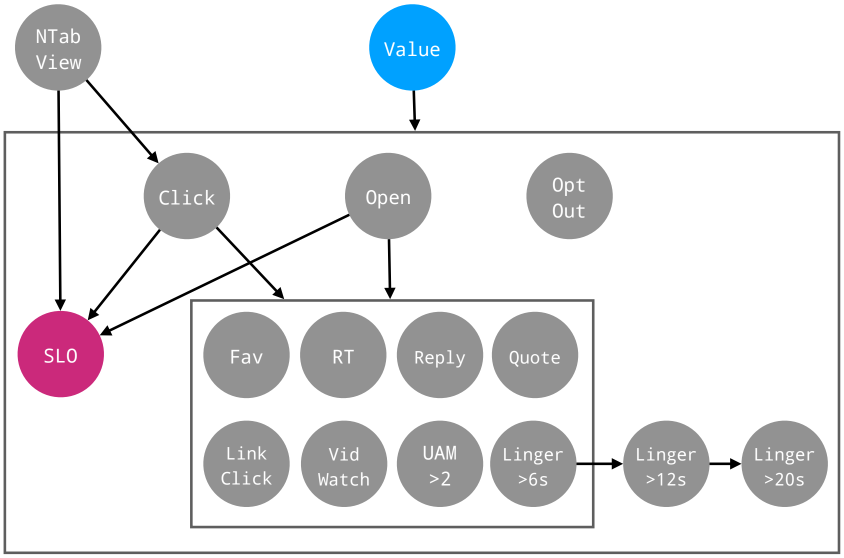


Figure 2: Bayesian network for Twitter notifications. An arrow from a node X to a box means that the node X is a parent of all the nodes in the box, e.g. `Click` and `Open` are parents of `Fav`, `RT`, ..., `Linger` $>$ `6s`. The latent variable `Value` is a parent of everything except `NTabView`. The measurement node `SLO` is highlighted in pink.

user can engage in many more interactions: they can favorite (`Fav`), retweet (`RT`), quote retweet (`Quote`), or reply (`Reply`) to the tweet; if the tweet has a link, they can click on it (`LinkClick`); if it has a video, they can watch it (`VidWatch`). In addition, other implicit signals are logged: whether the amount the user lingered on the tweet exceeds certain thresholds (`Linger` $>$ `6s`, `Linger` $>$ `12s`, `Linger` $>$ `20s`) and whether the number of user active minutes (`UAM`) spent in the app after clicking/opening the notification exceeds a threshold.

Furthermore, when the user is in the notification tab, the user can provide explicit feedback on a particular notification by clicking "See Less Often" (`SLO`) on it. Notably, unlike other types of behavior, the user does not need to actually click or open the notification before clicking SLO. However, we found empirically that users are more likely to click SLO after clicking or opening the notification, probably because they need to gain more information before making an assessment. Thus, in addition to `NTabView`, we also model `Click` and `Open` as parents of `SLO`.

Finally, at any time the user can opt-out of notifications to their phone home screen (`OptOut`). When the user decides to opt-out, it is attributed to any ML-based notification saw within a day of choosing to opt-out. Since ML-based notifications are relatively rare on Twitter (users usually get less than one a day), there are usually at most one or two notifications attributed to an opt-out event.

We model the latent variable V as being a parent of all behaviors except `NTabView` (whether or not the user saw the notification in their notifications tab or not). Since users may check their notifications tab for many other notifications, it is difficult to attribute `NTabView` to a particular notification, and so we consider it to be an exogenous, random event.

Identifying the joint distribution We fit our model on three days of data that contained of all user interactions with ML-based push notifications on Twitter. In Section 3, we proved that the target objective - the conditional distribution $\mathbb{P}(V = 1 \mid \mathbf{B}, A)$ - is uniquely identified from $\mathbb{P}(V = 1 \mid A = 1)$, $\mathbb{P}(V)$, $\mathbb{P}(\mathbf{B}, A)$, and $\mathbb{P}(\text{Pa}_{-V}(A) \mid A)$ (see Corollary 2). We set the four distributions as follows. We used `SLO` as our anchor variable A and assumed that $\mathbb{P}(V = 1 \mid A = 1) = 0$, i.e. a user never says “See less often” if they value the notification. The prior distribution of value $\mathbb{P}(V)$ was set to be uniform. The distribution of observed

$\mathbb{P}(V = 1 \mid \texttt{Behavior} = 1)$

`Behavior`	Naive Bayes	$\texttt{Click}, \texttt{Open} \nrightarrow \texttt{SLO}$	Full Model
`OptOut`	0	0	0
`Click`	0	0.316	0.652
`Open`	0	0.442	0.685
`UAM`	0	0.157	0.719
`VidWatch`	0	0.254	0.772
$\texttt{Linger} > \texttt{6s}$	0	0.264	0.802
`LinkClick`	0	0.320	0.836
`Reply`	0.358	0.570	0.932
$\texttt{Linger} > \texttt{12s}$	0	0.245	0.948
`Fav`	0.579	0.672	0.949
`RT`	0.680	0.720	0.956
$\texttt{Linger} > \texttt{20s}$	0.019	0.296	0.991
`Quote`	1.0	1.0	1.0

Table 1: The inferences made by LVMs with different DAGs. For each model and for each behavior, we list $\mathbb{P}(V = 1 \mid \texttt{Behavior} = 1)$ – how much evidence the model learns that a behavior provides for Value V (when all other behaviors are marginalized over).

behaviors $\mathbb{P}(\mathbf{B}, A)$ was set to the empirical distribution. The distribution $\mathbb{P}(\mathrm{Pa}_{-V}(A) \mid V)$ was estimated as described in Section 3.2 by using two sources of historical data, one in which notifications were sent at random and the other in which notifications were sent according to a recommendation algorithm.[8]

Evaluation of internal structure. Assessing our measure of "value" for validity will necessarily be an on-going and multi-faceted process. We do not, as typical of papers on recommendation, report engagement metrics. The reason is that if we expect our measure of "value" to differ from engagement, we cannot evaluate it by simply reporting engagement metrics. The evaluation of a measurement necessitates a more holistic approach. In Section 5, we describe the five categories of evidence for validity described by the *Standards for educational and psychological testing*, the handbook considered the gold standard on approaches to testing [AERA, APA, NCME, 2014].

Here, we focus on evaluating what is known as *evidence based on internal structure*, i.e whether expected theoretical relationships between the variables in the model hold. To justify why the structure of our Bayesian network is necessary, we compare our full model from Figure 2 to two other models: a naive Bayes model and the full model but without arrows from `Open` and `Click` to `SLO`. In Table 1, we show $\mathbb{P}(V = 1 \mid \texttt{Behavior} = 1)$ for all behaviors and models. As noted by prior work [Pearl, 2009, Halpern et al., 2016b], matrix adjustment methods can result in negative values when conditional independence assumptions are not satisfied. To address this, we clamp all inferences to the interval $[0, 1]$. We include the table of non-clamped inferences in the appendix (Table 2).

The first, simple theoretical relationship we expect to hold is that compared to observing no user interaction, observing any user behavior besides opt-out should increase the probability that the user values the tweet, i.e. $\mathbb{P}(V = 1 \mid \texttt{Behavior} = 1) < \mathbb{P}(V = 1) = 0.5$ for all $\texttt{Behavior} \neq \texttt{OptOut}$. Furthermore, we also expect some behaviors to provide stronger signals of value than others, e.g. that $\mathbb{P}(V = 1 \mid \texttt{Fav} = 1) > \mathbb{P}(V = 1 \mid \texttt{Click} =$

[8] We assume that the dataset of randomized notifications has a prior probability $\mathbb{P}_R(V = 1) = 0$ and the dataset of algorithmically chosen notifications has a prior probability $\mathbb{P}_C(V = 1) = 0.5$.

1).

The first model is the naive Bayes model, which simply assumes that all behaviors are conditionally independent given Value V. It does extremely poorly - almost all inferences have negative values and are clamped to zero, indicating that the conditional independence assumptions are unrealistic.

The second model is the full model except without arrows from `Click` and `Open` to `SLO`. It models all pre-requisite relationships between behaviors, i.e. if a behavior X is required for another behavior Y, then there is an arrow from X to Y. Compared to the naive Bayes model, the second model does not make mainly negative-valued inferences, indicating that its conditional independence assumptions are more realistic. However, relative to the prior, most behaviors actually reduce the probability of `Value`, rather than increase it!

After investigation, we realized that although users were not technically *required* to click or open the notification before clicking SLO, in practice, they were more likely to do so, probably because they needed to gain information before making an assessment. We found that explicitly modeling the connection, i.e. adding arrows from `Click` and `Open` to `SLO` was critical for making reasonable inferences. We believe this takeaway will apply across recommender systems. The user never has perfect information and may need to engage with an item before providing explicit feedback [Wen et al., 2019]. It is important to model the relationship between information-gaining behavior and explicit feedback in the Bayesian network.

Our full model satisfies the theoretical relationships we expect. All the behaviors that we expect to increase the probability of Value V do indeed do so. Furthermore, the relative strength of different types of behavior seems reasonable as well, e.g. $\mathbb{P}(V = 1 \mid \texttt{Fav} = 1)$ and $\mathbb{P}(V = 1 \mid \texttt{RT} = 1)$ are higher than $\mathbb{P}(V = 1 \mid \texttt{VidWatch} = 1)$ and $\mathbb{P}(V = 1 \mid \texttt{LinkClick} = 1)$.

The full model also makes more nuanced theoretical inferences. Recall that `UAM` is whether or not the user had high user active minutes after either clicking the notification from notifications tab or by opening the notification from their phone home screen. The model learns that `UAM` is a highly indicative signal after `Open`, but not after `Click`: $\mathbb{P}(V = 1 \mid \texttt{Open} = 1, \texttt{UAM} = 1) = 0.906$ and $\mathbb{P}(V = 1 \mid \texttt{Click} = 1, \texttt{UAM} = 1) = 0.641$. This makes sense because if the user clicks from notifications tab, it means they were already in the app, and it is difficult to attribute their high UAM to the notification in particular. On the other hand, if the user enters the app because of the notification, it is much more direct of an attribution.

It is clear that manually specifying the inferences our model makes would be very difficult. The advantage of our approach is that after specifying (a) the anchor variable and (b) the Bayesian network, we can automatically learn these inferences from data. Further, the model is able to learn complex inferences (e.g. that `UAM` is more reliable after `Open` than `Click`) that would be impossible to specify under the typical linear weighting of behaviors.

5 Assessing validity

Thus far, we have described our framework for designing a measure of "value", which can be used as a principled replacement for the ad-hoc objectives ordinarily used in engagement optimization. How do we evaluate such a measure? Notably, we do not advocate evaluating the measure purely through engagement metrics. If we expect our measure of "value" to differ from engagement, then we cannot evaluate it by simply reporting engagement metrics. Instead, the assessment of any measure is necessarily an ongoing, multi-faceted, and interdisciplinary process.

To complete the presentation of our framework, we now discuss approaches to assess the *validity* [Messick, 1987, AERA, APA, NCME, 2014, Reeves and Marbach-Ad, 2016] of a measurement. In the most recent (2014) edition of the *Standards for educational and psychological testing*, the handbook considered the gold standard on approaches to testing, there are five categories of evidence for validity [AERA, APA, NCME, 2014]. We visit each in turn, and describe how they translate to the recommender system setting, using Twitter as an example.

Evidence based on content refers to whether the content of a measurement is sufficient to fully capture the target construct. For example, we may question whether a measure of "socio-economic status" that includes income, but does not account for wealth, accurately captures the content of the construct [Jacobs and Wallach, 2019]. In the recommender engine setting, content-based evidence asks us to reflect on whether the behaviors available on the platform are sufficient to capture a worthy notion of the construct "value". For example, if the only behavior observed on the platform were clicks by the user, then we may be skeptical of any measurement of "value" derived from user behavior. What content-based evidence makes clear is that to measure any worthy notion of "value", it is essential to design platforms in which users are empowered with richer channels of feedback. Otherwise, no measurement derived from user behavior will accurately capture the construct.

Evidence based on cognitive processes. Measurements derived from human behavior are often based on implicit assumptions about the cognitive processes subjects engage in. Cognitive process evidence refers to evidence about such assumptions, often derived from explicit studies with subjects. For example, consider a reading comprehension test. We assume that high-scoring students succeed by using critical reading skills, rather than a superficial heuristic like picking the answers with the longest length. To gain evidence about whether this assumption holds, we might, for instance, ask students to take the test while verbalizing what they are thinking.

Similarly, in the recommender engine setting, we want to verify whether user behaviors occur for the reasons we think they do. On Twitter, one might think to use `Favorite` as an anchor for Value V, assuming that $\mathbb{P}(V = 1 \mid \texttt{Favorite} = 1) \approx 1$. However, users actually favorite items for reasons that may not reflect value – like to bookmark a tweet or to stop a conversation. Cognitive process evidence highlights the importance of user research in assessing the validity of any measure of "value".

Evidence based on internal structure refers to whether the observations the measurement is derived from conform to expected, theoretical relationships. For example, for a test with questions which we expect to be of increasing difficulty, we would assess whether students actually perform worse on later questions, compared to earlier ones. In the recommender system context, we may have expectations on which types of user behaviors should provide stronger signal for value. In Section 4, we evaluated internal structure by comparing $\mathbb{P}(V = 1 \mid \texttt{Behavior} = 1)$ for all behaviors.

Evidence based on relations with other variables is concerned with the relationships between the measurement and other variables that are external to the measurement. The external variables could be variables which the measurement is expected to be similar to or predict, as well as variables which the measurement is expected to differ from. For example, a new measure of depression should correlate with other, existing measures of depression, but correlate less with measures of other disorders. In the recommender system context, we might look at whether our derived measurement of "value" is predictive of answers that users give in explicit surveys about content they value. We could also verify that our measure of "value" does not differ based on protected attributes, like the sex or race of the author of the content.

Evidence based on consequences. Finally, the consequences of a measurement cannot be separated from its validity. Consider a test to measure student mathematical ability. The test is used to sort students into beginner or advanced classes with the hypothesis that all students will do better after sorted into their appropriate class. If it turns out that students sorted by the test do *not* perform better, that may give us reason to reassess the original test. In the recommender system context, if we find that after using our measurement of value to optimize recommendations, more users complain or quit the platform, then we would have reason to revise our measurement.

6 Summary

We have presented a framework for designing an objective function that captures a desired notion of "value". In line with the principles of measurement theory, we treat "value" as a theoretical construct which must be operationalized. Our framework allows the designer to operationalize "value" in a principled manner by specifying only an *anchor variable* and the structure of the Bayesian network. Through these two choices, the designer has the flexibility to give "value" subtly different meanings.

We applied our approach on the Twitter platform on millions of users. We do not, as typical of papers on recommendation, report engagement metrics. The reason is that if we expect our measure of "value" to differ from engagement, we cannot evaluate it simply by reporting engagement metrics. Instead, we discussed established ways to assess the validity of a measurement and how they translate to the recommendation system setting. For the scope of this work, we focused on assessing *evidence based on internal structure* and found that our measure of "value" satisfied many desired theoretical relationships.

Acknowledgements

We thank Naz Erkan for giving us the opportunity and freedom to conduct this work through her bold leadership and savvy managerial support. We thank Prakhar Biyani for his extensive effort in helping us apply our approach at scale at Twitter. We thank Tom Everitt for feedback on a draft of the paper.

References

Michael D Ekstrand and Martijn C Willemsen. Behaviorism is not enough: better recommendations through listening to users. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 221–224, 2016.

Hongyi Wen, Longqi Yang, and Deborah Estrin. Leveraging post-click feedback for content recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 278–286, 2019.

David J Hand. *Measurement theory and practice: The world through quantification.* Arnold London, 2004.

Simon Jackman. Measurement. In *The Oxford Handbook of Political Methodology*, chapter 9. Oxford University Press, 09 2009. ISBN 9780199286546.

Judea Pearl. *Causality.* Cambridge university press, 2009.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing. *Standards for educational and psychological testing.* AERA, 2014.

Samuel Messick. Validity. *ETS Research Report Series*, 1987(2):i–208, 1987.

Todd D Reeves and Gili Marbach-Ad. Contemporary test validity in theory and practice: A primer for discipline-based education researchers. *CBE—Life Sciences Education*, 15(1):rm1, 2016.

Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. *arXiv preprint arXiv:1912.05511*, 2019.

Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 1–10. IEEE, 2012.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288, 2013.

Yoni Halpern, Steven Horng, Youngduck Choi, and David Sontag. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association*, 23(4): 731–740, 2016a.

Yoni Halpern, Steven Horng, and David Sontag. Clinical tagging with joint probabilistic models. In *Conference on Machine Learning for Health Care*, 2016b.

Judea Pearl. On measurement bias in causal inference. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 425–432. AUAI Press, 2010.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Kenneth J Rothman, Sander Greenland, and Timothy L Lash. *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008.

Appendices

$\mathbb{P}(V = 1 \mid \texttt{Behavior} = 1)$

Behavior	Naive Bayes	$\texttt{Click}, \texttt{Open} \nrightarrow \texttt{SLO}$	Full Model
OptOut	-99.74	-0.932	-0.072
Click	-1.194	0.316	0.652
Open	-0.366	0.442	0.685
UAM	-1.092	0.157	0.719
VidWatch	-0.475	0.254	0.772
Linger $>$ 6s	-0.525	0.264	0.802
LinkClick	-0.302	0.320	0.836
Reply	0.358	0.570	0.932
Linger $>$ 12s	-0.254	0.245	0.948
Fav	0.579	0.672	0.949
RT	0.680	0.720	0.956
Linger $>$ 20s	0.019	0.296	0.991
Quote	1.0	1.0	1.0

Table 2: The same inferences as in Table 1, except without clamping to $[0, 1]$.