2018

# Using Latent Variable Models to Improve Causal Estimation

Huseyin Oktay

# USING LATENT VARIABLE MODELS TO IMPROVE CAUSAL ESTIMATION

A Dissertation Presented

by

HÜSEYİN OKTAY

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2018

College of Information and Computer Sciences

# USING LATENT VARIABLE MODELS TO IMPROVE CAUSAL ESTIMATION

A Dissertation Presented

by

HÜSEYİN OKTAY

Approved as to style and content by:

David D. Jensen, Chair

Ramesh K. Sitaraman, Member

Benjamin M. Marlin, Member

Krista J. Gile, Member

James Allan, Chair<br>
College of Information and Computer Sciences

*To Zeynep and the kids, the sources of joy in my life.*

*Be a rainbow in someone else’s cloud.*

*Maya Angelou*

# ACKNOWLEDGMENTS

This thesis immensely benefited from its committee chair, David Jensen. I felt privileged to have David on my side at every stage during my graduate studies. He patiently and continuously nudged me to become a curious researcher while showing me not only the ways of overcoming difficulties but also the joys of making scientific contributions. His support—both professionally and personally—has been out of this world. I sincerely thank him for all he has done for me—simple and plain.

I am grateful for my other committee members Krista Gile, Benjamin Marlin, Ramesh Sitaraman for their valuable comments and feedback. I would like to thank Dan Corkill for his general research advice and for accompanying me during my first business trip at UMass; Cindy Loiselle for tirelessly and meticulously editing early drafts of my manuscripts (coming from a non-native writer, reading those must have felt very close to being tortured); Leanne Leclerc for keeping me on track with all the degree requirements; Deborah Bergeron for smoothly taking care off all the administrative paperwork; Oliver Brock for his support and mentorship in my early years in the graduate program.

My academic family in the Knowledge Discovery Laboratory with its past and present members made my experience during graduate school fulfilling and fun. I would like to extend my special thanks to David Arbour, Andrew Fast, Lisa Friedland, Dan Garant, Amanda Gentzel, Michael Hay, Philip Kirlin, Marc Maier, Katerina Marazopoulou, Jennifer Neville, Matt Rattigan, and Brian Taylor for their friendship and colleagueship. I would like to thank other members of the department—Elif Aktolga, Ethem Can, Laura Dietz, Sam Huston, İbrahim Uysal, Zeki Yalnız—for all the stimulating discussions as well as fun and useful exchanges.

I would like to extend my special thanks to my managers at Apple—Aravind Gowrisankar, Abbey McGrew, Michael Sorensen, Tal Tversky—for their flexibility while I work on the last remaining piece for my dissertation.

I am thankful to my parents on both sides for giving me the freedom and unconditional support to pursue my goals of studying abroad. I am especially grateful my mother-in-law, Ayşe Ertem, for all the sacrifices she has done for my family by helping take care of the princesses, which spared valuable time for me to work on my dissertation.

Last but not the least, I would like to extend my special thanks to my beautiful and loving wife, Zeynep. This thesis would not exist without her enduring support. Her continuous and unwavering encouragement enabled me to reach the finish line. I am truly grateful for her profoundly enriching my life with joy and happiness. I am so lucky to have found her.

# ABSTRACT

## USING LATENT VARIABLE MODELS TO IMPROVE CAUSAL ESTIMATION

FEBRUARY 2018

HÜSEYİN OKTAY

B.Eng., BOĞAZİÇİ UNIVERSITY

M.Sc., UNIVERSITY OF MASSACHUSETTS, AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor David D. Jensen

Estimating the causal effect of a treatment from data has been a key goal for a large number of studies in many domains. Traditionally, researchers use carefully designed randomized experiments for causal inference. However, such experiments can not only be costly in terms of time and money but also infeasible for some causal questions. To overcome these challenges, causal estimation methods from observational data have been developed by researchers from diverse disciplines and increasingly studies using such methods account for a large share in empirical work. Such growing interest has also brought together two arguably separate fields: machine learning and causal estimation, and this thesis also contributes to this intersection.

Specifically, in observational data researchers have lack of control over the data generation process. This results in a fundamental challenge: the presence of confounder variables (i.e., variables that affect both treatment and outcome). Such variables, when not adjusted statistically, can result in biased causal estimates. When confounder variables are observed, many methods can be used to adjust for their effect. However, in most real world observational data sets, accurately measuring all potential confounder variables is far from feasible, hence important confounder variables are likely to remain unobserved. The central idea of this thesis is to explicitly account for unobserved confounders by inferring their values using a predictive model.

This thesis presents three main contributions in the intersection of machine learning and causal estimation. First, we present one of the earliest application of causal estimation methods from social sciences to social media platforms to answer three causal questions. Second, we present a novel generative model for estimating ordinal variables with distant supervision. We also apply this model to data from US Twitter user population and discover variation in behavior among users from different age groups. Third, we characterize the behavior of an effect restoration model based on graphical models with theoretical analysis and simulation studies. We also apply this effect restoration model with predictive models to account for unobserved confounder variables.

# TABLE OF CONTENTS

**Page**

**CHAPTER**

# LIST OF TABLES

**Table** **Page**

# LIST OF FIGURES

**Figure** **Page**

# CHAPTER 1

# INTRODUCTION

Estimating the causal effect of a treatment from data has been a key goal for a large number of studies in many domains. Examples include medical studies estimating the effect of a medication on patients' health; public policy studies estimating the effect of a social program on the welfare of a certain group of people; and business studies estimating the effect of an online marketing campaign on customers' shopping behavior.

Traditionally, researchers use carefully designed randomized experiments for causal inference [28, 93]. However, such experiments can be costly in terms of time and money [59]—recruitment of subjects and recording their outcomes throughout experiments may take time; furthermore, often times there are a large number of experiments required to rule out alternative explanations. More importantly, there may be causal effects of interest in which randomization of treatment might be unethical (such as randomly assigning subjects to a smoking group to estimate its effect on lung cancer) or might be impossible (such as randomly assigning gender to estimate its effect on drug use). To overcome these limitations of randomized experiments, causal estimation methods from observational data have been developed by researchers in diverse disciplines and increasingly account for a large share of studies in empirical work [4, 10, 37].

Such growing interest has also brought together two arguably separate fields: machine learning and causal estimation. Several recent studies in this intersection are particularly notable. Athey et al. [10] survey many recent causal estimation methods for policy decisions and discuss how recent developments in machine learning field can help to develop new causal estimation methods [9, 10]; Wager et al. [104] develop a causal random forest method to estimate heterogeneous treatment effects; Varian [102] discusses how machine learning methods can model the counterfactual for a treatment and showcase a study about the effect of online advertisement on increasing sales; on a complementary study, Kleinberg et al. [41] argue that not all policy problems are *causal* and some are *predictive* and contend that machine learning can help effectively solve prediction policy problems. This thesis also contributes to the mutual interaction between machine learning and causal estimation from observational data.

The key difference of observational data from experimental data is that the details of the underlying data generation process—especially the treatment assignment mechanism—are unobserved or uncertain and not controlled by the researcher [36]. This lack of control over the data generation process in observational data creates a fundamental challenge: the presence of confounder variables (i.e., variables that affect both treatment and outcome). Such variables, when not adjusted for statistically, can result in biased estimates of causal effect [4].

Consider estimating the effect of two different medication options for treating a certain disease, one affordable and one costly option. A naïve analysis using observational data might compare well-beings of patients using one medication to that of patients using the other medication. Furthermore, such a hypothetical analysis might credit the difference in the outcome to the difference in the medication used. This process could result in an unsound causal estimate as it ignores potential confounder effects. For example, socioeconomic status of patients might be correlated with medications they can afford as well as their general state of well-being. The patients who use the affordable medication might also be in poor general health condition. Conversely, patients who use the costly medication might be in relatively good general health condition. Taking the naïve difference in health condition might include the effect due to socioeconomic condition (i.e., the confounding effect) along with the direct effect of the medication (i.e., treatment effect). Hence, directly comparing the outcomes of populations (i.e., without accounting for confounders) might provide biased estimates [94, 111].

When confounder variables are observed, researchers have developed many statistical methods to account for their effect, and these methods are sometimes referred to as quasi-experimental designs (QEDs). A few examples of such designs are covariate adjustment [35, 94], propensity score matching [5, 63, 89], and instrumental variables [4, 7]. These methods have been applied and extended by investigators from diverse disciplines due to the availability of large and rich observational data sets capturing detailed interactions in diverse domains [40, 50, 105]. Just to highlight a few recent studies, Aral et al. [7] estimate the peer influence on running behavior using an instrumental variable design; Peysakhovich et al. [79] develop a method that combines observational data with experimental data resulting in decreased number of experiments required to reliably estimate a causal effect (i.e., efficient experimentation); Krishnan and Sitaraman [45] identify the effects of video streaming quality on viewer engagement in content delivery networks using a matching design; Bornfeld et al. [18], identify the effect of newly introduced badges on user behavior in three Stack Exchange websites, using a natural experiment design.

In most real world observational data sets, capturing all potential confounder variables is far from being feasible, hence important confounder variables might be unobserved. When unobserved confounder variables exist, causal estimation from observational data is even more challenging and, as discussed earlier, can result in biased estimates. However, the increasing availability of large and rich data sets suggests that proxy variables for potential unobserved confounders can be inferred from other observed and correlated variables.

The central idea of this thesis is to explicitly account for unobserved confounders with effect restoration by inferring their values using predictive models. First, we employ predictive models to estimate the values of confounder variables. Second, we use such inferred values as the proxy variables of unobserved confounders with one caveat that the proxy variables are measured with error. We use an effect restoration model based on graphical models as our measurement error model to deal with this estimation error, and adjust for the confounding effect due to unobserved variables. We propose this mechanism as a novel method to remove bias in causal estimation due to unobserved confounder variables.

Although the ideas presented in this thesis are generally applicable to a wide range of domains, the focus of this thesis is mostly causal inference in social media domains due to two main reasons. First, social media platforms provide a framework in which social phenomena can emerge and be measured on an unprecedented scale, breadth, and depth [50, 105]. For example, Twitter has several hundreds of millions of daily active users around the globe, and their post activities and relationships are recorded over time resulting in a large and rich observational data set.

Second, computational social scientists increasingly use data from social media platforms to answer causal questions (e.g., [6, 68, 79]). However, often times social scientists want to adjust for potential confounding effects of traditional demographic variables such as age and ethnicity and such variables are almost always unobserved on data from such platforms. For example, estimating the effects of using social media on TV consumption, many social scientists think that age can be a confounding variable (i.e., younger users both tend to overwhelmingly use social media and tend to substantially consume TV). If we use social media data to answer such a question without adjusting for the effects of confounding variables, we might overestimate its effect on TV consumption potentially resulting in harsh policies in restricting social media access. Here, we enrich arguably popular data from social media platforms with latent variable models to adjust for unobserved confounder variables.

Third, several platforms make their data either publicly available (e.g., the Stack Exchange websites), or provide ways to obtain samples of their data through public firehouses, partnerships, sales, or challenge contests (e.g., Twitter, Yelp, or Netflix) enabling replications of research results and follow-up studies by other researchers.

The contributions of this thesis include:

- *An early application of causal estimation methods based on QEDs to social media platforms*—We demonstrate one of the earliest use of QEDs to data about social media platforms. We apply three different QEDs to answer causal questions about social media systems, specifically the Stack Overflow website. First, we use a matching design to estimate the effect of having a high-quality answer on the number of subsequent answers. Our results suggest no significant effect of having a high-quality answer on the subsequent posts. Second, we use an interrupted time-series design to estimate the effect of a specific badge, the *epic* badge, on user engagement on the website. Our results suggest that engagement is sustained until the badge is received, but is reduced after. Third, we use a natural experiment design to identify the effect of answer ordering on the number of up-votes an answer gets. Our results find no significant effect suggesting that users ignore ordering while voting.

- *A novel generative model for estimating ordinal variables with distant supervision*—We develop a novel generative model to estimate ordinal variables with distant supervision. Specifically, we use this model to estimate age using first names and evaluate our model using voter registration data. We then apply our method to understand the demographic breakdown of users on Twitter and find that 18-29-year-old user group is the largest among Twitter users. We show that our method can eliminate limitations of other methods such as surveys performed by Pew Research in estimating the demographic breakdown of social media users resulting in complete visibility of all age groups in user populations. We also perform analysis to estimate different usage patterns of different age groups on Twitter in terms of their topical interests and follow relationships. We find that follow relationships show strong evidence of assortative mixing for young and senior users (e.g., young users follow other young users) but weak evidence of assortative mixing for middle-aged users.

- *An effect-restoration mechanism based on graphical models that improves causal estimation by accounting for unobserved confounders*—We characterize the behavior of an effect restoration model based on graphical models with theoretical analysis and simulation studies. First, we empirically confirm prior work showing that the effect restoration adjustment reduces bias only when the variable measured with error is a confounding variable. We also show that the relative benefit of effect restoration is the highest for estimating small treatment effects with large confounding bias. By using a real world data set from a randomized experiment, we show that the effect restoration adjustment removes bias more than its natural alternatives. Second, by leveraging graphical models, and d-separation, we show, for the first time, that simple rules and typical temporal ordering assumptions are sufficient to identify whether a variable measured with error is a confounding variable. This knowledge determines if using effect restoration for that variable can reduce bias. Finally, through simulation studies, we show that the effect restoration adjustment can reduce bias for an unobserved confounding variable, when estimates for that variable are available from an independent process, such as a predictive model, along with the corresponding error distribution.

The organizational structure of the remainder of this document is as follows. Chapter 2 summarizes the necessary background for this thesis, including social media platforms, graphical models and their causal extensions, d-separation, and the challenges of causal estimation. This chapter also describes the problem statement of this thesis. Chapter 3 presents one of the earliest application of QEDs to observational data from the Stack Overflow website. Chapter 4 develops a novel generative model to estimate ordinal variables with distant supervision and shows a specific case to estimate age using first names. This chapter also illustrates the results of its application to US Twitter user base estimating different types of social behavior on the platform. Chapter 5 characterizes the use of the graphical model-based effect restoration mechanism to deal with measurement error in confounding variables through theoretical analysis and simulation studies. This chapter also applies the effect restoration mechanism along with independent estimation methods to adjust for unobserved confounder variables. Chapter 6 concludes the thesis and suggests future directions.

# CHAPTER 2

# BACKGROUND AND PROBLEM STATEMENT

This chapter reviews several key underlying concepts used in this thesis. First, we identify the unique opportunities and challenges of using data from social media platforms. Second, we review the relevant concepts in graphical models with a focus on their use in representing causal knowledge. Third, we discuss the challenges of causal estimation from observational data sets. Finally, we state the main problem addressed in this thesis.

## 2.1 Social Media Platforms: Opportunities and Challenges

Since the beginning of the 21$^{st}$ century, the use of social media platforms have increased tremendously, enabling them to digitally capture many social interactions between users such as friendship, communication, and financial exchange [40, 50, 105]. As an unintended side effect, they have substantially increased the measurement capabilities of social scientists studying general human behavior. They have enabled broader data collection (by observing different interactions of user behavior), deeper data collection (by observing such interactions at the transaction level), and larger-scale data collection (by observing millions and for some platforms billions of users all together) [40, 50, 105]. For example, Facebook[1] has more than 2 billion monthly active users and can capture friendship, communication, and personal relationships among its users [19, 103]. Such big and rich data sets, coupled with advancements in computational tools to analyze them [2, 102], provide new opportunities for researchers to study social phenomena that are analogous to the opportunities created when researchers started to use microscopes in the 1600's [40, 50].

[1] www.facebook.com

However, these opportunities come with many challenges as well. Preserving and assuring the privacy and security of users' data on these platforms are valid concerns [50]. Within the context of academic studies, these social systems, by providing new ways of interactions, also alter the human behavior creating feedback loops for behaviors under study [90]. Furthermore, many algorithmic features on these platforms interact with each other and changes in one feature may have unforeseen effects on other parts of the platform [49]. Finally, the user base in these platforms might have population bias (e.g., younger users tend to use Twitter [27, 68]) and unobserved important variables, making the generalization of research results challenging to the entire population.

In this thesis, we aim to address the problem of unobserved variables by leveraging the advancements in machine learning models and the richness of big data. We propose that one can estimate proxy variables for important unobserved variables, using predictive models that rely on other related observed fields.

## 2.2 Bayesian Networks

A Bayesian network is a widely used graphical model to capture joint probability distributions among variables [30, 39, 44]. The structure of a Bayesian network is a directed acyclic graph (DAG) defined by a set of vertices and a set of edges, $G =< V, E >$. For example, Figure 2.1 represents a Bayesian network with $V = \{A, B, C, D, E\}$ and $E = \{< A, B >, < A, C >, < B, C >, < B, D >\}$.

Each vertex, $v \in V$, represents a random variable. Each edge, $e \in E$ represents a probabilistic dependency between variables forming the edge $e =< A, B >$. Given a directed edge, $A \rightarrow B$, vertex $A$ is called a parent of vertex $B$. Vertex $B$ is called a child of vertex $A$. A vertex $v_d$ is called a descendant of vertex $v_i$, if there is a *directed* path from $v_i$ to $v_d$. For example, in Figure 2.1, vertex $D$ is one of the descendants of vertex $A$. A non-descendant vertex of $v_i$ is, simply, a vertex that is not a descendant. For example, vertex $A$ is a non-descendant for vertex $D$, in Figure 2.1.

Bayesian networks, by definition, satisfy the local Markov property.

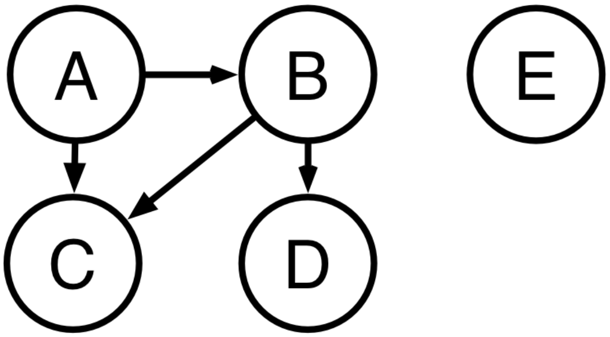


Figure 2.1: Example Bayesian network

**Definition 2.2.1. Local Markov property :** Given a causal graph $G = (V, E)$ , a variable $X \in V$ is independent of every other variable except $X$'s descendants given its parent variables.

For example in the Bayesian network in Figure 2.1, the Local Markov property implies: $D \perp\!\!\!\perp \{A, C, E\} \mid B$.

More generally, Bayesian networks provide a compact framework to encode (in)dependencies in a given domain between pairs of variables. *d-separation* is a graph-based criterion to identify conditional independence relationships from the structure of a Bayesian network. *d-separation* criterion conceptually links statistical independence relationships among variables with the connectedness of variables in networks.

Here we review the *d-separation* criterion to provide the necessary background for understanding contributions in this thesis as oppose to explaining it thoroughly. Before, giving the definition of *d-separation*, let us provide some useful definitions related to the concept of blocking of paths, using the Bayesian network in Figure 2.1, as our example.

**Definition 2.2.2.** Given the following path between vertex $A$ and $D$, $A \rightarrow B \rightarrow D$, conditioning on $B$ would **block** the path.

**Definition 2.2.3.** Given the path between vertex $A$ and $B$, $A \rightarrow C \leftarrow B$, conditioning on $C$ would **unblock** the path.

**Definition 2.2.4.** Vertex $v_i$ and $v_j$ are **d-connected**, if there is an unblocked path between them.

**Definition 2.2.5.** Vertex $v_i$ and $v_j$ are **d-separated**, if they are not d-connected.

For example, for the Bayesian network in Figure 2.1, by using d-separation criterion, we can devise the following relationships among the variables (by no means the exhaustive list of dependencies).

- $E \perp\!\!\!\perp \{A, B, C, D\}$
- $A \perp\!\!\!\perp D \mid B$
- $A \not\!\perp\!\!\!\perp D \mid C$
- $A \perp\!\!\!\perp D \mid \{C, B\}$
- $C \perp\!\!\!\perp D \mid B$
- $C \not\!\perp\!\!\!\perp D \mid A$

Different causal Bayesian networks on the same set of variables can imply the same set of conditional independence relationships, and this is formally defined as Markov equivalence.

**Definition 2.2.6** (Markov Equivalence)**.** Let two DAGs, be $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ on the same set of nodes $V$. $G_1$ and $G_2$ are called *Markov equivalent* if and only if, based on the Markov condition, they entail the same conditional independencies and dependencies.

For example, in Figure 2.2, in row a, both of the graphical models imply the same dependence relationship between X and Y (i.e., $X \not\!\perp\!\!\!\perp Y$), and hence they are Markov equivalent.

**Definition 2.2.7** (Markov Equivalence Class)**.** A Markov equivalence class, $\boldsymbol{\zeta}$, is a set of DAGs, where all pairs of $(G_i, G_j)$, such that $G_i, G_j \in \boldsymbol{\zeta}$ and $G_i = (V, E_i)$,, $G_j = (V, E_j)$, are Markov equivalent.

## 2.3 Causal Bayesian Networks

Bayesian Networks, with the following additional assumption, have been extended to represent causal dependencies [73, 97].

**Assumption 1** (Causal Markov Assumption)**.** *Given a causal graph $G = (V, E)$ , a variable $X \in V$ is independent of every other variable except $X$'s effects conditional on all of its direct causes.*

For example, a causal interpretation of the Bayesian network in Figure 2.1 implies that A is a cause of B, and B is an effect of A. A and D are causally independent given B.

Pearl [73] introduces *interventions* and the *do-operator* to formalize causal estimation using Causal Bayesian networks. In this framework, interventions imply actively setting the values of a variable instead of passively observing them. The *do-operator* distinguishes interventions from mere observations in Bayesian networks. $P(Y|do(X = x^{'}))$ implies the interventional distribution for $Y$ when the value of $X$ is set to $x^{'}$. However, $P(Y\,|X = x^{'})$ implies the conditional distribution for $Y$ when the value of $X$ is passively observed as $x^{'}$. From a graphical representation perspective, $P(Y|do(X = x^{'}))$ changes the structure of the graph by removing the incoming edges to $X$, whereas $P(Y \mid X = x^{'})$ implies no change in the structure, just mere observation of $X$.

Given the causal semantic of intervention, the average treatment effect of a binary $X$ on $Y$ (i.e., $ATE$) can be calculated as:

$$ATE = E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)] \tag{2.1}$$

where $E[Y|do(X = x)]$ implies the expected value of the interventional distribution $P(Y|do(X = x))$.

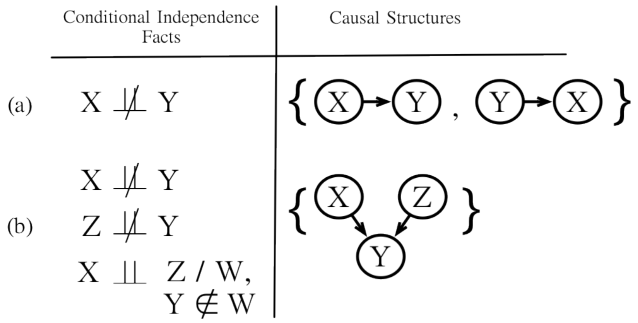


Figure 2.2: Conditional independence facts and the corresponding causal structures that can explain such facts are enumerated. For the first example (first row), two potential causal structures listed are statistically indistinguishable. However, for the second example (second row), the causal structure is uniquely identified given the facts.

## 2.4 Challenges of Causal Estimation

### 2.4.1 Correlation Underdetermines Causality

One of the main challenges of causal estimation is that association underdetermines causality. In other words, different causal structures can explain the observed set of conditional independence facts in the data. This defines the limits of causal discovery algorithms on how much they can learn from data.

More formally, several causal discovery algorithms that aim to learn the underlying structure of a causal Bayesian network from data make the following two key additional assumptions.

**Assumption 2** (Causal Sufficiency)**.** *All common cause variables of all variables represented in the causal Bayesian network are observed.*

**Assumption 3** (Faithfulness)**.** *All the independence relationships implied by the causal Bayesian network, $G$, are present in any population sample, $P$, causally represented by such network.*

With these assumptions, a Markov equivalence class for the observed conditional independence relationships in a given data set can be determined. For example, an association between variables X and Y can be explained by at least two different causal models, as shown in Figure 2.2, even when all the important variables are measured (i.e., assuming no latent variables). These two models are statistically indistinguishable based on the data.

For example, watching excessive amounts of TV (i.e., X) may be correlated with performing violent behavior (i.e., Y). A person is more likely to engage in violent behavior if he watches excessive TV. However, ignoring any potential common cause variables for the sake of argument, this correlation might be either because of watching excessive TV makes people violent (i.e., X $\rightarrow$ Y) or conversely, because violent people extensively watch TV (i.e., Y $\rightarrow$ X). With only the knowledge about correlation, we cannot further infer the cause and the effect.

However, with another variable and the corresponding set of conditional independence facts, the underlying causal structure can be uniquely identified, as shown in Figure 2.2 on the second row. In more detail, when X is marginally correlated with Y, and similarly Z is marginally correlated with Y, if X and Z are conditionally independent (i.e., conditionally not correlated) given a set **W** that does *not* include Y, a unique causal structure that satisfies all these constraints, as shown in Figure 2.2 (i.e., a Markov equivalence class of size 1).

In general, one of the main challenges of causal discovery is that there may be multiple statistically indistinguishable models (i.e., models that are Markov equivalent) that can explain the observed conditional independence facts, even when all variables are observed. It is important to distinguish among those models since each of these models implies a different set of actions to take to change the desired outcome.

### 2.4.2 Latent Common Cause Variables

The size of the set of statistically indistinguishable causal models that can explain a set of observed conditional independence facts is even larger if several important variables are latent (i.e., unobserved). Specifically, when latent variables could exist (when causal sufficiency assumption is lifted), the set of statistically indistinguishable models that can explain the association between X and Y includes at least three more causal structures, as shown in Figure 2.3.

For example, building on the earlier example, watching excessive amounts of TV (i.e., X) and performing violent behavior (i.e., Y) might be correlated. However, a certain personality trait might be a common cause for both watching TV and performing violent behavior (i.e., latent common

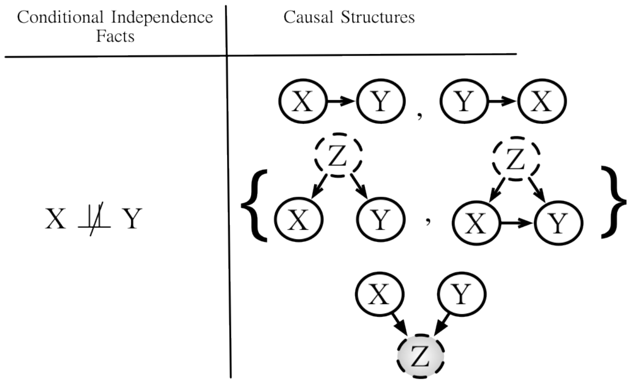


Figure 2.3: Conditional independence facts and causal structures with latent variables that can explain such facts are enumerated. Dashed nodes represent latent variables, and a shaded node represent conditioning on the corresponding variable.

cause, Z). Hence, hypothetically, all the observed correlation can be explained by a latent common cause represented as the third possible structure in Figure 2.3. Alternatively, when a latent common cause exists, there can still be a direct effect as shown in the fourth structure.

Finally, the fifth structure corresponds to a case with selection bias [99] where a common effect of X and Y exists and is conditioned on. For example, arguably, people watching excessive amounts of TV might have limited number of friends in their circles. Similarly, people performing violent behavior might have limited number of friends in their circles. A study performed on people with limited number of friends might result in a spurious correlation between watching excessive TV and performing violent behavior. Careful randomized experimentation and propensity score matching are some ways proposed in the literature to deal with sample selection bias [14, 93].

### 2.4.3 Faithfulness

The faithfulness assumption, stated earlier, is a complement to causal Markov assumption that the implied dependencies by $G$ should be expected in $P$. If this does not hold, the causal dependencies in $G$ might be missing in $P$, making learning $G$ from $P$ challenging. If two paths in $G$ with opposite effects perfectly cancel each other out, the resulting population sample may not be faithful to the underlying structure.

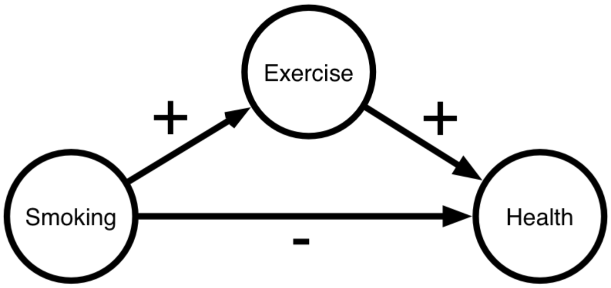


Figure 2.4: In this structure, smoking has two competing effects on health: a negative direct effect, and a positive indirect effect. When competing effects perfectly cancel each other out, faithfulness assumption is violated.

For example, in the hypothetical world represented in Figure 2.4, smoking negatively affects health, and positively affects exercise. Exercise eventually positively affects health. If the direct negative effect of smoking on exercise cancels out with its indirect positive effect on health through exercise, then the resulting population would not be faithful to the underlying graphical model structure, making the learning of causal structures from data challenging.

### 2.4.4 Unbiased and Consistent estimates

The main goal for causal inference is to estimate a treatment effect. Let's assume $TE$ is the true treatment effect, and $TE_{est}$ is our estimator. $TE_{est}$ is an *unbiased* estimator for $TE$, if the expected value of $TE_{est}$ is $TE$. Mathematically:

$$TE = \mathbf{E}[TE_{est}(x_1, x_2, ..., x_n)].$$

Carefully designed randomized experiments guarantee unbiased estimates of the treatment effect [8, 54, 74].

Furthermore, $TE_{est}$ is a *consistent* estimate of $TE$ if:

$$TE = \lim_{n \to \infty} TE_{est}(x_1, x_2, ..., x_n)$$

where $TE_{est}(x_1, x_2, ..., x_n)$ implies the value of the estimator using $n$ instances. A consistent estimate is asymptotically unbiased and most of the causal estimation methods from observational data guarantee a consistent estimate of a treatment effect [74, 87, 93, 97].

## 2.5 Problem Statement

The focus of this thesis is causal estimation when latent common cause variables exist. The main idea is to use external statistical processes to explicitly infer the values of latent variables, and then use those inferred values for causal estimation.

Many of the existing causal estimation methods assume causal sufficiency, the assumption that all common cause variables are known and accurately measured [55, 56, 62, 73, 96, 97, 113]. However, in many real world cases, latent variables might exist [13, 25, 97], and ignoring those latent common causes might introduce a bias in causal estimation [94, 111]. Spirtes et. al. [98] lift causal sufficiency assumption and identify sufficient conditions based on joint conditional dependencies to determine causal structures from data when latent variables might exist. Based on these conditions authors propose the FCI algorithm, Zhang [111] extends the FCI algorithm by providing additional edge orientation rules and shows that the extended algorithm is complete, in the sense that the edge orientation rules cover all possible cases to identify causal structures for any given joint dependency set, though set of uniquely identifiable causal structures are limited resulting in large number of DAGs in Markov equivalence classes. Rattigan and Jensen [85] introduce relational blocking as an operator for causal discovery algorithms that can leverage the relational structure of rich data sets. For causal estimation, instrumental variable designs have been shown to adjust for both observed and unobserved latent variables [4, 58], though the challenge lies in finding an instrument for a treatment variable.

The proposed approach in this thesis deals with latent common cause variables in a distinct and novel way.

First, the method explicitly infers the values of latent variables by exploiting other external sources of information. Such sources can be predictive models for latent variables using other

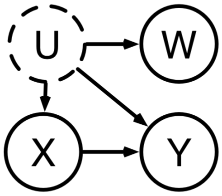


Figure 2.5: The latent variable, U, can also be the cause of the other observed signal, W. However, conditioning on the observed signal does not necessarily eliminate the bias due to the latent variable (i.e., block the path between X and Y that goes through the latent variable).

information. For example, when age information about people is a latent variable but information about their first names is available, we might use a statistical model to predict the age of those people using their first names. We expect that the availability of large data sets in unprecedented breadth, depth, and scale [9, 50, 105] have the potential to present abundant opportunities to predict relevant latent variables, using other correlated information.

Second, the proposed approach suggests conditioning on estimated values of latent variables to reduce bias in causal estimation. One might suggest to directly condition on the other information in the data as a proxy of the latent common cause variable while performing causal estimation. In our example above, the suggestion might be to use first name values directly in causal estimation rather than inferring latent age values through a statistical model.

We suggest several challenges with this approach. First, such a variable may have many possible values such that in a data set with a reasonable size, each possible value might have limited number of actual instances. For example, more than 150,000 different first name values are registered in the US Social Security data about baby names, and conditioning merely on names is likely to provide subgroups with a handful of instances. Second, if a latent variable is causal for the other observed information, as shown in Figure 2.5, then conditioning on the observed signal does not necessarily eliminate the bias due to the latent variable (i.e., conditioning on W does not block the path between X and Y that goes through the latent variable) [31, 47].

# CHAPTER 3

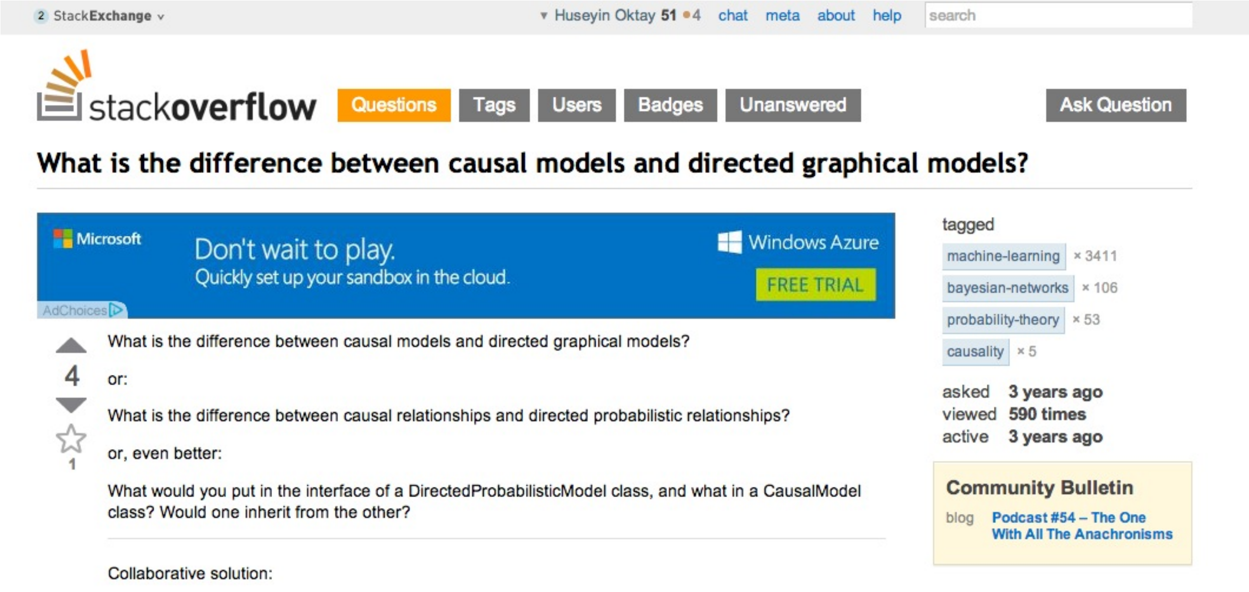
# CAUSAL ESTIMATION USING QUASI-EXPERIMENTAL DESIGNS

In this chapter, we report one of the earliest causal analysis of an arguably popular social media platform using quasi-experimental designs (QEDs)[1]. Traditionally, social scientists use QEDs for causal estimation from observational data. Here, we present results from one of the earliest applications of such designs to causal questions about social media platforms. Specifically, we briefly describe three different QEDs and apply them to answer causal questions related to the Stack Overflow website[2], estimating cause-and-effect relationships about this question and answer platform. We then discuss the assumptions and limitations of QEDs specifically about threats to validity when latent common cause variables exist. This chapter provides the motivations to deal with such latent variables for unbiased causal effect estimations.

This chapter is organized as follows. First, we describe the question and answer platform used in our study, the Stack Overflow website. Second, we review the related literature about social media platforms and knowledge sharing platforms with a specific focus on causal analysis both before and after our work. Third, we provide a brief overview of QEDs. Fourth, we present our analysis of applying three distinct QEDs to Stack Overflow estimating relevant relationships on the website. Finally, we summarize our conclusions and highlight the limitations of QEDs with respect to unobserved variables.

[1] Much of the content of this chapter is derived from: H. Oktay, B. J. Taylor, and D. Jensen. (2010).

[2] http://www.stackoverflow.com



What is the difference between causal models and directed graphical models?

What is the difference between causal models and directed graphical models?

or:

What is the difference between causal relationships and directed probabilistic relationships?

or, even better:

What would you put in the interface of a DirectedProbabilisticModel class, and what in a CausalModel class? Would one inherit from the other?

Collaborative solution:

Figure 3.1: A screenshot of an example question on the Stack Overflow website.

## 3.1 The Stack Overflow Website

Stack Overflow is one of the more than 100 online platforms built using Stack Exchange [3], an online framework for constructing sites in which users can exchange knowledge through questions and answers. Each of these platforms has a specific topical focus and an extensive amount of rich data capturing interactions among users. Specifically, Stack Overflow focuses on questions related to programming and has over 7 million registered users, more than 10 million monthly visits, along with almost 36 million posts. A screenshot of an example question on Stack Overflow is shown in Figure 3.1.

There are five main entities in Stack Overflow, as shown in a simple entity-relationship diagram in Figure 3.2: (1) users, (2) posts, which represent both questions and answers in the system, (3) comments, (4) votes, and (5) badges. There are three main actions in Stack Overflow. First, users can ask questions related to programming. Second, users can share their knowledge by providing an answer to a particular question. Third, users can vote up or down questions and answers that

[3] http://stackexchange.com/sites

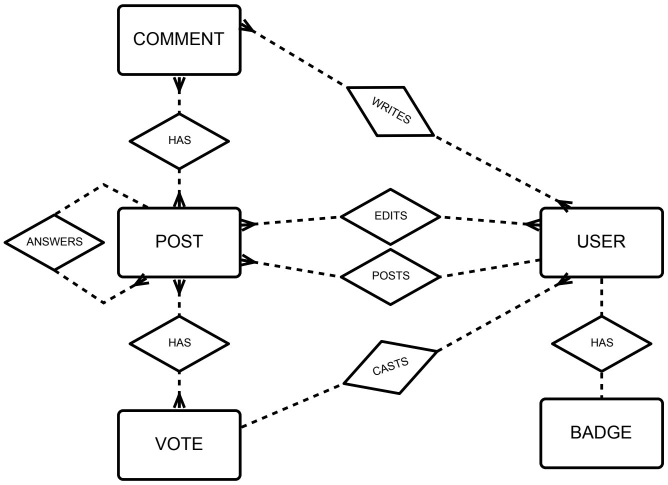


Figure 3.2: Entities and relationships in Stack Overflow.

they like or dislike. As users take these actions, they both contribute to the platform and gain reputation points as well as badges.

Understanding the interactions among different entities by identifying cause-and-effect relationships has valuable benefits not only while managing the day-to-day operations of the system but also while studying the user behavior in knowledge sharing platforms. The data for the Stack Overflow website (along with all other Stack Exchange websites) is publicly available providing a desirable opportunity for research.

## 3.2 Related Work

Social media platforms, where collective user behavior emerges and can be measured in detail, provide an unprecedented research opportunity [50, 101, 105]. Many researchers from diverse fields including computer science and social sciences have performed studies using data from such platforms. We review this diverse related literature in three distinct categories: (1) Predictive models and macro modeling of user dynamics; (2) Causal studies with experimental and non-experimental data; (3) Studies focusing on question-and-answer platforms and the Stack Overflow website.

As for predictive and macro-modeling studies, Bakshy et al. [11] studied the role of social networks on content adoption using the social network in Second Life,[4] a multiplayer game in a virtual world. Lerman et al. [51] studied the role of social networks in promoting content via voting mechanisms on the web using data from a social news aggregator website (i.e., Digg[5]). In another study, Lerman et al. [52] developed a predictive model for social media content popularity based on user behavior, again using data from Digg. Ratkiewicz et al. [84] proposed a model to explain the macro dynamics of online popularity by using Wikipedia entries and web pages. Wilkinson [106] described a model about macro user-behavior to explain the contributions in online peer production systems using data about Wikipedia.

As for causal studies, Kohavi et al. [42] published a practical guide to controlled randomized experiments on the Web, summarizing their lessons learned while developing the online experimentation platform for Microsoft[6]. Salganik et al. [91] performed an experimental study in an artificial online cultural market to identify the effect of social influence on inequality and unpredictability of success. As for non-experimental data, Aral, Muchnik, Sundararajan [5] used dynamic propensity score matching on an instant messaging platform to disentangle the effect of homophily from peer influence on product adoption, using the underlying social network based on instant messaging interactions.

As for studies about question and answer platforms, Raban et al. [81] discussed motivations behind user contribution in peer-to-peer knowledge sharing systems using qualitative analysis. Zhang et al. [112] found that the expertise networks in question and answer platforms are structurally different than other online networks such as the World Wide Web. Adamic et al. [1] identified that in Yahoo! Answers, interactions in certain categories mirror interactions in expertise sharing forums; however, certain other categories that involve discussions and posts about everyday advice mirror interactions in social networks. Furthermore, certain users chose to contribute with a narrow focus

[4] http://secondlife.com

[5] http://digg.com

[6] http://exp-platform.com/experiments-at-microsoft/

in categories, and certain users contribute with a diverse focus across categories. Finally, they used these insights to develop a model to predict whether a particular answer would be the best answer for a given question. Kumar et al. [46] provided a theoretical model for the evolution of question and answer platforms over time capturing the two different actions of users (asking vs answering questions) as a two-sided market. They, then, used Yahoo! Answers and Stack Overflow as two case studies to gather empirical evidence for their theoretical model.

In our study, described in this chapter, we provide one of the earliest studies [7] about peer-to-peer knowledge sharing platforms with a specific focus on causal inference. Specifically, we use non-experimental data from Stack Overflow and employ QEDs for causal inference. Our work was cited by a number of subsequent causal studies about social media.

For example, Krishnan and Sitaraman [45] identified the effects of video streaming quality on viewer behavior in content delivery networks using a matching design. Reis et al. [86] used matching methods on Twitter to identify the effect of exercise on mental health. Sharma et al. [95] use instrumental variable design to identify the causal effect of online recommender systems on the viewer traffic of product pages. Kusmierczyk et al. [48] used instrumental variable design to identify the effect of first-time badges (i.e., badges awarded for the first occurrence of a particular type of action) on user activity on the Stack Overflow website. Bornfeld et al. [18], using a natural experiment design, identified the effect of newly introduced badges on user behavior in three separate Stack Exchange websites.

## 3.3 Quasi-Experimental Designs

In an ideal experimental scenario, random assignment of experimental units to values of treatment is used for obtaining unbiased estimates of causal effects [28, 29]. One of the key advantages of randomly assigning treatment is that it eliminates the confounding bias due to both observed and latent variables [73]. We illustrate random assignment using graphical model terminology in Fig-

[7] https://meta.stackexchange.com/questions/134495/academic-papers-using-stack-exchange-data

ure 3.3a. When the values of a treatment variable are assigned completely randomly, as depicted by variable $R$, all incoming edges to a treatment variable are removed resulting in an unbiased estimate of the causal effect (i.e., no back door path from treatment to outcome).

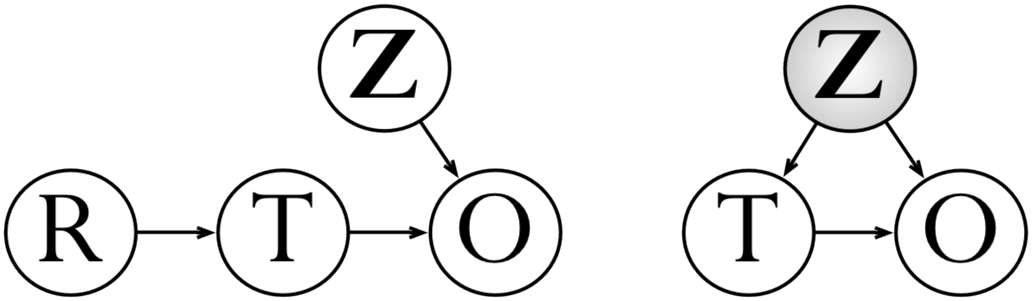
In many social media platforms, randomized experimentation may be unavailable to researchers due to economic and experimental integrity concerns [6, 45, 69]. It may be too costly to design and deploy experiments over millions of customers. Additionally, platform managers may be meticulous about the design of experiments and hesitant to deploy any experiment that might sour users' experience on the platform.

When random assignment of treatment is either impossible or infeasible, QEDs [20, 93] are widely used. Other than lacking random assignment, QEDs have purposes and characteristics similar to those of randomized experiments. Designs generally work by identifying an experimental unit that has undergone treatment and comparing it to another experimental unit that has not undergone treatment but that is similar to the corresponding treatment unit, in almost all other aspects.

Traditionally, these methods have been used and developed by social and biological scientists to answer policy questions. However, in the past decade, increased availability of rich and large observational data through social media platforms and wearable devices provided new application domains of existing designs as well as new opportunities for the development of new methods [6, 7, 45, 50, 69]. In this chapter, we present results from one of the earliest applications of three specific QEDs to estimate cause-and-effect relationships about a social media platform, Stack Overflow.

## 3.4 Using Designs to Discover Causal Knowledge

The Stack Overflow website is a platform where users interact with each other through questions and answers related to programming. The managers of the platform might have a strategic goal to include all relevant content on the platform. Since there are no curators for content, the platform relies on users to both ask and answer questions. To achieve this goal, managers might want to discover the factors that change the engagement levels on the platform, such as features



(a) Graphical model of random assignment.

(b) Graphical model of matching.

Figure 3.3: Randomized assignment of treatment and the matching design are represented in graphical model formalism, respectively. Note that when treatment (T) is randomized (R), there cannot be any other cause for (T) other than (R), resulting in statistical control of potential common cause variables. Whereas, the matching design attempts to model the effect of potential common causes (**Z**), to obtain a consistent causal estimate.

that increase the number of posts from users. This requires performing causal analysis and the rich observational data that is already being collected for operational purposes can also be used for that purpose. Specifically, we present our results using several matching designs, an interrupted time series design, and a natural experiment design to identify causal dependence within the Stack Overflow platform.

### 3.4.1 The Matching Design

Since observational data generally do not have the random assignment of treatment variables, matching designs are used to avoid confounding bias by conditioning on observed covariates in each matched pair. Using graphical model formalism, we illustrate the difference between a randomized experimental design and a matching design in Figure 3.3b. The matching design identifies pairs of units where one of the units has received a treatment and the other has not such that those units are similar in *all* other observed measures. Existing matching methods use different similarity metrics and weighting mechanisms to assign units to control and treatment groups such as propensity scores and Euclidean distance of covariate vectors [36, 89, 100]. The validity of the causal conclusions drawn from a matching design improves as the matched pairs become more similar to each other. A causal question related to Stack Overflow that we can attempt to answer with this design is:

> *Does posting a high-quality answer for a particular question cause a reduction in the number of subsequent answers posted by other users?*

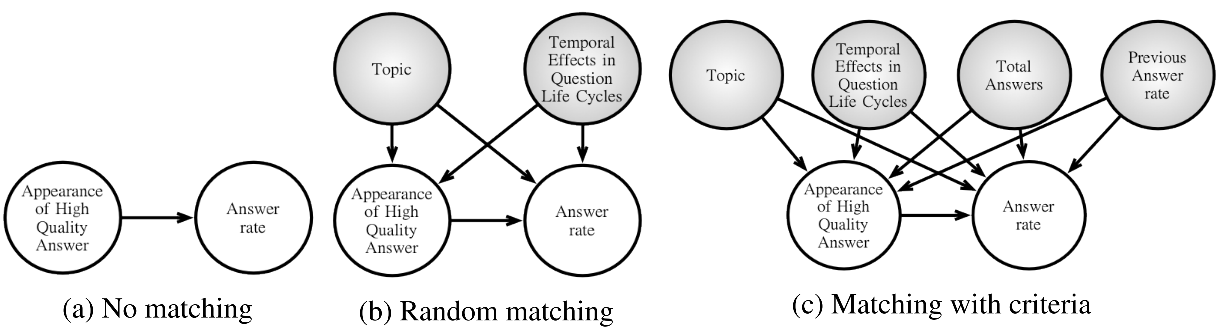
The question has direct implications for website policies related to maximizing user contribution. For example, if the answer to the question is a negative effect, a policy that publishes the answers after a certain total number of answers are reached (or a certain amount of time passed) might result in higher number of posts and can be desirable for website managers.

To answer this question, first, we define a method for determining a high-quality answer. We use a key characteristic of the Stack Overflow site: the user who asks a question can select one of the answers as the *accepted answer*. We assume the accepted answer is a high-quality answer for a particular question. The accepted answer is often selected long after it is initially posted. Because of this time lag, we can examine the effect of quality on the number of subsequent answers.

To illustrate the importance of matching criteria, we apply, in progression, three different versions of the matching design [69] as summarized in Figure 3.4 using graphical models. In the first design, we only examine questions in the treatment group (i.e., questions with accepted answers) with no matching, as shown in Figure 3.4a. In the second design, we have treatment-control pairs where control questions are randomly selected from questions that have the same tag. Tags are labels for questions to relate them to corresponding topics or concepts (e.g., machine-learning, causality). In the third design, we match treatment and comparison questions by requiring a much stronger similarity within pairs based on their answer rates.

The outcome is the change in answer rate $\Delta t$ minutes before and after the treatment is applied. The first design is a simple statistical analysis that evaluates the change in answer rate before and after an accepted answer occurs. We show the results of this analysis in the *No Matching* column of Table 3.1. We find a negative change in answer rate for different $\Delta t$ values.

This suggests that the answer rate is greater before the treatment and that there is a decrease in the number of answers provided after a high-quality answer is posted. However, it is unclear whether this change is caused by the appearance of the eventually accepted answer (i.e., treatment).


Appearance of High Quality Answer
Answer rate
Topic
Temporal Effects in Question Life Cycles
Appearance of High Quality Answer
Answer rate
Topic
Temporal Effects in Question Life Cycles
Total Answers
Previous Answer rate
Appearance of High Quality Answer
Answer rate

(a) No matching (b) Random matching (c) Matching with criteria

Figure 3.4: Graphical models corresponding to the different matching cases. Shaded variables represent variables that are being conditioned. The goal of the analysis is to estimate the causal effect of a high-quality answer on the subsequent change in answer rate.

The decrease in the answer rate could be caused by the intrinsic change in the answer rate due to temporal effects in question life-cycle rather than the presence of the accepted answer. For example, questions might get high exposure right after they are initially posted because they are featured on the homepage. This might result in a large number of answers initially, and as time passes, the exposure fades away resulting in a small number of answers.

To eliminate the temporal effects in question life cycles, we can use a basic matching design as shown in Figure 3.4b. We pair each treatment question with a random control question to better compare the difference in behavior. The idea is both questions in the matched pair go through a similar life cycle, hence we can adjust for its effect. We randomly select questions that have the same tag to adjust for the variability due to topical differences. The outcome measure for this design is the difference between the answer rate change of a treatment question (i.e., $T_{arc}$) and the answer rate change of its comparison question (i.e., $C_{arc}$) within the matched pair.

As shown in the *Random Pairs* column of Table 3.1, with this version of the design that also conditions on temporal effects, we conclude that at least for several $\Delta t$ values the difference between the treatment and comparison questions is insignificant when we compare the random-pair design to no matching. We conclude, at least in those cases, that the accepted answer has no effect on the subsequent answer rate. We can also observe that for $\Delta t$ values where there appears to be an effect, the size of the effect is smaller than that estimated using the no matching design. This

| Time | Experiments | | |
|---|---|---|---|
| $\Delta t$<br>(in minutes) | No<br>Matching | Random<br>Pairs | Matching |
| 15 | -0.78 | -0.66 | NS |
| 20 | -0.45 | NS | NS |
| 25 | -0.52 | -0.25 | NS |
| 30 | -0.36 | -0.24 | 0.18 |
| 60 | -0.24 | -0.12 | NS |
| 90 | -0.10 | -0.07 | NS |
| 120 | -0.10 | NS | NS |
| 150 | -0.03 | NS | NS |
| 180 | -0.05 | NS | NS |

Table 3.1: Our analysis indicate no significant effect of having a high quality answer on answer rate. For *No Matching* experiments, differences in answer rate for each time interval are shown. For *Random Pairs* and *Matching*, differences between the answer rate change for the treatment group and the answer rate change for the control group are shown. NS means *Not Significant*. Values are number of answers per hour.

shows evidence that we can have a more thorough analysis by using designs that can adjust for potential common cause variables.

Although this design matches a pair of questions, we are not guaranteed to find highly similar pairs. For example, answer rate before treatment can be an important variable that can affect the high-quality answer as well as the answer rate after treatment, as shown in Figure 3.4c. In such cases, any difference we observe may be partially or fully due to the inherent difference in their previous answer rate rather than the high-quality answer provided.

To create better-matched pairs in the third design, we combine two criteria. First, we require the treatment question and the control question to have nearly the same number of overall answers provided. Second, we want the matched pair to have a similar previous answer rate before treatment (for the specified time interval, from $[t - \Delta t, t]$ where $t$ is the time the high-quality answer is provided for a treatment question). For example, in Figure 3.5 we illustrate the total number of answers for three different questions over time; the dashed vertical line shows the time the accepted answer is posted for question I. According to the criteria specified, for question I, question II

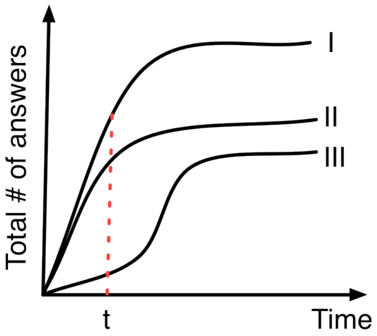


Figure 3.5: The total number of answers over time for three different hypothetical questions. The vertical dashed line represents the time that the accepted answer is posted for question I (i.e., a question in the treatment group).

is a closer match than question III because II has not only similar number of total answers but also similar answer rate with question I before treatment. Using the criteria described above, we matched 200 pairs of questions with the same tag.

This design uses the same outcome measure as the second design and the right-most column of Table 3.1 shows the results obtained with this design. By also conditioning on previous answer rate and a total number of answers, the final version of the matching design indicates that having a high-quality answer has no significant effect on answer rate for almost all $\Delta t$ values, whereas the previous designs do show a statistically significant effect. Even for some $\Delta t$ values where the results of the random pair design (i.e., the second design) suggest that a high-quality answer has an effect, the final matching design results suggest the effect is not significant.

More generally, in an ideal matching design, the matching criteria should eliminate the effects of all possible confounders as possible explanations for the observed effect. Depending on the causal question, a researcher can identify matching criteria to adjust for almost all the variables that can simultaneously influence both the treatment and the observed effect, such as the topics of the questions, activities of users, and extrinsic rewards provided to users by the site (e.g., their badges or reputation points). However, additional possible confounders could be latent, and these variables might invalidate causal conclusions based on matching designs. Regardless of the matching criteria used, researchers should continue to consider alternative explanations involving latent variables at the end of the analysis and look for ways of inferring those potential latent variables, if they suspect that any exist.

### 3.4.2 The Interrupted Time-Series Design

In the *interrupted time series* design, we observe an outcome variable of a unit during a certain time interval, $\Delta t$, that includes an interruption for another variable (i.e., the treatment variable) [20, 93]. This observation of the same unit over $\Delta t$ lets us adjust for latent variables within the unit and thus rule out some threats to validity. For example, one causal question in Stack Overflow that can be examined using this design is:

> *For users, does receiving the epic badge cause an increase in their posting activity?*

For Stack Overflow managers, a strategic goal for the platform might be to keep high levels of user engagement at all times. Badges can provide motivation for users to continue their contributions by providing recognition and distinction among other users. Understanding if and how much a badge changes user behavior is an important causal question.

For example, the *epic* badge is given to users who earn 200 reputation points (the daily maximum) 50 times. For this design, the treatment is receiving the epic badge, and the outcome is the number of posts. We normalize the number of posts for each user by taking the difference between the average number of posts by that particular user before and after treatment over a 60-day interval. To be more precise, for each user we calculate two average values: (1) the average number of posts before treatment, (2) the average number of posts after treatment. We normalize the number of posts for each user with the corresponding average value [8]. There are 54 users with the epic badge in our data set, and they obtain the badge at different physical times. By relatively aligning such different physical time points that users obtain the badge, the effect of other exogenous events

[8] This is a within subject design where for each user her own behavior before treatment serves as a control subject.

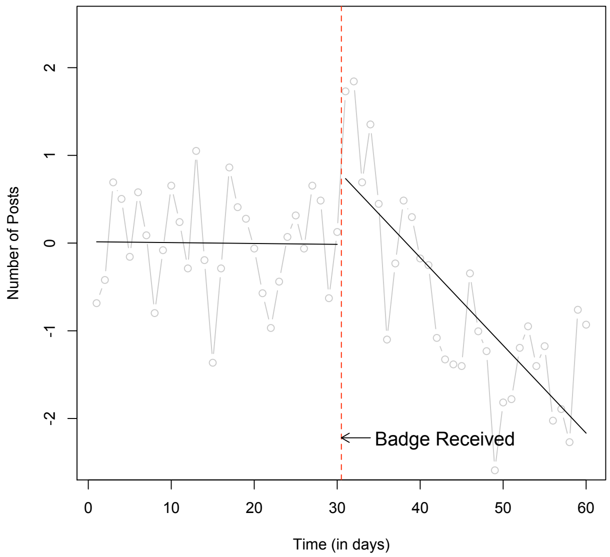


Figure 3.6: We report that the average number of posts by users that obtain the epic badge decreases after receiving the badge. Values are normalized for each user by subtracting the corresponding average number of posts before and after treatment.

that might occur on the site (e.g., being featured on a popular blog post or executing a marketing campaign) are assumed to be averaged out.

In Figure 3.6, we show the results of the interrupted time series design. The vertical dashed line represents the 30-day mark for each user at which the epic badge was received. We plot the average normalized number of posts for each day on the y-axis. The first linear model is for the average number of posts before the badge is received and the second linear model is for the average number of posts after the badge is received. The slope of the first line is $-0.001$, and this slope is not significantly different than 0 ($p = 0.94$). The slope of the second line is $-0.10$, and this slope is significantly different than 0 ($p < 0.01$). We observe a significant negative slope after the badge is received. Our results suggest that obtaining the epic badge reduces the number of posts created by the corresponding users. An alternative badge mechanism where tiered-badges with increasing level of exclusivity can help sustain contributions from users.

### 3.4.3 The Natural Experiment Design

A natural experiment is a condition within an observed data set that approximates the conditions of a randomized experiment, particularly randomized assignment of treatment. Such a condition can occur if a social media system changes a single aspect, such as a user interface, and has data collected both before and after the change. While the system change was never intended to be a treatment used in an experiment, the data can be analyzed as if it was.

A causal question in Stack Overflow for this design is:

> *Does the order in which answers are displayed within the Stack Overflow interface cause the number of votes received by each of those answers?*

The ordering of answers has been a subject of public debate on the Stack Overflow website. Some users argued that the previous policy (which used the chronological order of posting) was unfair to users who provide quality answers at a later time[9]. Clearly, this public discussion (which featured hundreds of posts and thousands of votes about the issue over several years) suggests that users and moderators cared about the perceived effects of answer order.

The causal question can be examined by using data generated by the eventual policy change in Stack Overflow. In Stack Overflow, answers for a question are sorted in descending order in terms of their net number of votes. To break ties, two different approaches were used in the system. Before August 2009, ties were broken in terms of the creation date of the answers. Older posts got higher priority and were listed higher on the page when there was a tie in the number of votes. After August 2009, Stack Overflow managers changed their policy and decided to break ties randomly, which removed bias in answer ordering that favored older posts. Answers are now sorted randomly when they have received the same net number of votes.

To benefit from this versatile natural experiment, we consider tie-breaking votes before and after the policy change as our instances. A tie-breaking vote is a vote-up that is cast for an answer that is in a tie with exactly one other answer. Our treatment variable is the way the answers are ordered in a tie situation (i.e., either from oldest to newest or randomly). Our outcome variable

[9] See, particularly: https://meta.stackexchange.com/questions/9731/fastest-gun-in-the-west-problem

| Vote Number | Number of Instances | P-Value | $\chi^2$-Statistic | Frequency |
|---|---|---|---|---|
| 1 | 87405 | $< 2.2e^{-16}$ | 2684.2 | 0.41 |
| 2 | 45899 | $< 2.2e^{-16}$ | 2208.0 | 0.39 |
| 3 | 24908 | $< 2.2e^{-16}$ | 1042.6 | 0.40 |
| 4 | 14260 | $< 2.2e^{-16}$ | 807.8 | 0.38 |
| 5 | 8555 | $< 2.2e^{-16}$ | 532.8 | 0.38 |
| 6-8 | 12291 | $< 2.2e^{-16}$ | 638.3 | 0.39 |
| 9-12 | 6434 | $< 2.2e^{-16}$ | 337.7 | 0.38 |
| 13-17 | 3748 | $< 2.2e^{-16}$ | 145.3 | 0.40 |
| 18-26 | 3593 | $< 2.2e^{-16}$ | 109.4 | 0.41 |
| 26-665 | 9728 | $< 2.2e^{-16}$ | 160.6 | 0.44 |

Table 3.2: The results of a chi-square test on the frequency of votes for the older answer before the policy change. Degree of freedom is 1 for each stratum.

is the frequency of voting for the older answer when there is a tie. If the order of answers has an effect on voting, we would expect to see a significant difference between the frequency of voting for the older answer before and after the policy change.

We randomly choose more than 200,000 tie-breaking votes, both before and after the policy change. For each vote, we assign a binary value that is $0$ if the vote is for the newer answer in the tie situation or $1$ if the vote is for the older answer. Then we do a chi-square test to determine if those values are significantly different than a binomial distribution where the success probability is $0.5$. In such a binomial distribution, we would expect to have equal numbers of votes for newer answers and for older answers. We do the same test for the votes before and after the policy change.

The frequency of tie-breaking votes for the older answer is $0.40$ before the policy change, and the chi-square test reveals that this frequency is significantly different than 0.5 with $\chi^2 = 8487.76$ and $p < 0.001$. Similarly, after the policy change, the frequency of tie-breaking votes for the older answer is $0.40$, and the chi-square test reveals that this frequency is also significantly different than 0.5 with a $\chi^2 = 8424.14$ and $p < 0.001$. These results show that users are more likely to vote for the newer answer than for the older answer regardless of the policy change. The results could also imply that the newer answers are often of a higher quality than older answers.

| Vote Number | Number of Instances | P-Value | $\chi^2$-Statistic | Frequency |
|---|---:|---|---:|---:|
| 1 | $87436$ | $< 2.2e^{-16}$ | $2652.1$ | $0.41$ |
| 2 | $46079$ | $< 2.2e^{-16}$ | $1915.5$ | $0.40$ |
| 3 | $24756$ | $< 2.2e^{-16}$ | $1253.2$ | $0.39$ |
| 4 | $13954$ | $< 2.2e^{-16}$ | $884.9$ | $0.37$ |
| 5 | $8718$ | $< 2.2e^{-16}$ | $527.3$ | $0.38$ |
| 6-8 | $12323$ | $< 2.2e^{-16}$ | $670.7$ | $0.38$ |
| 9-12 | $6530$ | $< 2.2e^{-16}$ | $271.7$ | $0.40$ |
| 13-17 | $3970$ | $< 2.2e^{-16}$ | $118.5$ | $0.41$ |
| 18-26 | $3668$ | $< 2.2e^{-16}$ | $139.7$ | $0.40$ |
| 26-665 | $9810$ | $< 2.2e^{-16}$ | $177.5$ | $0.43$ |

Table 3.3: The results of a chi-square test on the frequency of votes for the older answer after the policy change. Degree of freedom is 1 for each stratum.

To assess the effect of the policy change, we also compare the frequency of tie-breaking votes for the older answer before and after the policy change. We do a two-sample unpaired t-test with two-tail analysis, and the results reveal no significant difference between the frequencies of tie-breaking votes for the older answer before the change and after the change with $t = 0.35$, degrees of freedom = $433640$, and $p = 0.73$. The data does not support the hypothesis that the policy change had an effect on voting behavior.

Recall that in our dataset of tie-breaking votes, we have votes that correspond to a vote-up for an answer for which the pre-tie-breaking vote count (i.e., the number of existing vote-ups of an answer right before the tie-breaking vote) is in a tie situation with exactly one other answer. This vote-up in our data set breaks this tie between these two answers (i.e., answer pairs) either by voting for the older answer or the newer answer. A closer observation of this vote distribution for the tie-breaking votes in our dataset is shown in Figure 3.7. In these figures, the x-axis shows the number of pre-tie-breaking votes and the y-axis shows the frequency of tie-breaking votes. For example, from Figure 3.7, we can see that in our dataset there are more than 20,000 tie-breaking votes that break the tie between answer pairs for which the pre-tie-breaking vote count is three. Both answers in the pair already have three vote-ups, and the tie- breaking vote in our data set

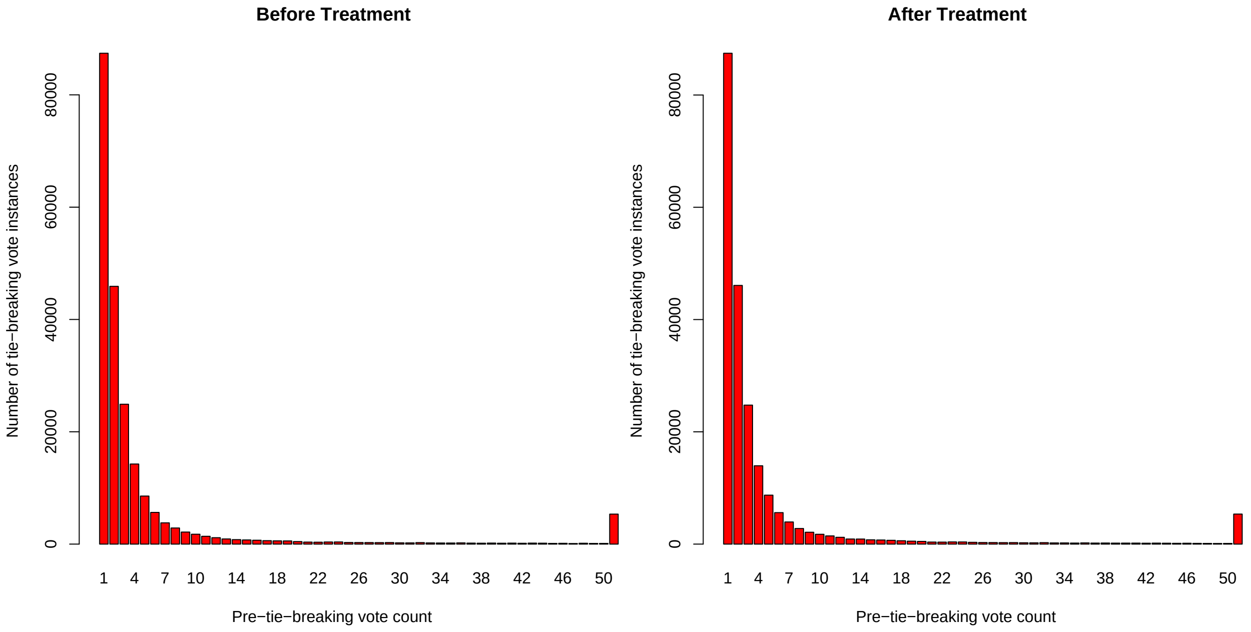


Figure 3.7: Histogram for tie-breaking votes before and after the policy change. The last bar is the aggregate of all the tie-breaking votes of answers for which the pre-tie-breaking vote count is greater than 50.

breaks this tie by increasing the vote number for one of the answers (either the older answer or the newer answer) by one.

The histograms in Figure 3.7 show that many of the tie- breaking votes are cast for answer pairs when there is exactly one pre-tie-breaking vote for each answer in these pairs. The number of qualifying-vote instances decreases rapidly as the number of pre-tie-breaking votes goes up for two reasons. First, since vote-ups are counted cumulatively for an answer, there are more vote-up instances that correspond to answers for which the pre-tie-breaking vote counts are small than instances that correspond to answers for which the pre-tie-breaking vote counts are large. Second, being in a tie situation with another answer when the pre-tie-breaking vote count is small is more likely than when the pre-tie-breaking vote count is large. For these two reasons, we get more tie-breaking votes that correspond to answers for which the pre-tie-breaking vote counts are small when we randomly sample from vote-up instances (i.e., there is a bias towards small pre-tie-

| Vote Number | P-Value | $t$-Statistic | Frequency Before | Frequency After |
|---|---|---|---|---|
| 1 | 0.81 | 0.23 | 0.41 | 0.41 |
| 2 | 0.02 | 2.32 | 0.39 | 0.40 |
| 3 | 0.02 | −2.34 | 0.40 | 0.39 |
| 4 | 0.25 | −1.15 | 0.38 | 0.37 |
| 5 | 0.90 | 0.12 | 0.38 | 0.38 |
| 6-8 | 0.66 | −0.43 | 0.39 | 0.38 |
| 9-12 | 0.15 | 1.44 | 0.38 | 0.40 |
| 13-17 | 0.19 | 1.31 | 0.40 | 0.42 |
| 18-26 | 0.41 | −0.82 | 0.41 | 0.40 |
| 26-665 | 0.69 | −0.40 | 0.44 | 0.43 |

Table 3.4: The results of a two-sampled unpaired t-test with two-tail analysis for comparing the frequency of voting for the older answer before and after the policy change.

breaking vote counts for data instances both before and after the treatment as shown in Figure 3.7).

To better account for the difference in the distributions seen in Figure 3.7, we stratify our data points into 10 strata, as shown in the vote number column of Table 3.2. For each stratum, we perform the chi-square test to see if the frequency of tie-breaking votes for older answers is different than the binomial distribution where the probability is 0.5. Tables 3.2 and 3.3 show the results for the data points in each stratum before and after the policy change, respectively. Those tables show that, for each stratum, the frequency is significantly different than 0.5, both before and after the policy change, suggesting that regardless of the existing vote count, users vote-up for the recently posted answer more frequently than expected by random chance.

We also performed a two-sample unpaired t-test with two-tail analysis to see whether, in each stratum, the frequency of voting for the older answer before the policy change is different than the frequency after the policy change. Table 3.4 summarizes the results of these tests. For a generous $p$-value threshold $0.05$, except for strata 2 and 3, the results were not significant, indicating that we cannot accept the hypothesis that the frequency of the vote for the older answer differs before and after the policy change. For strata 2 and 3, although we get a significant result, the difference in the frequency values is extremely small, suggesting a very small effect, if any.

These type of policy changes that enable useful analysis for impactful questions might be more frequent than generally expected. For example, another recent policy change in the Stack Overflow system sets up another natural experiment to assess whether there is a causal relationship between reputation points and asking questions. Users get reputation points when their questions or answers get a vote-up. Before the recent policy change, users received 10 points for a vote-up, both for a question and an answer. However, after this change, users get 5 reputation points after a vote-up for their questions and still get 10 reputation points after a vote-up for their answers. The goal of this change is to decrease the number of questions a user asks and increase the number of answers that a user provides. This natural experiment can be leveraged to identify the effect of the change in vote-up reputation points on user behavior. The treatment variable is the change in vote-up reward for questions. The outcome variable is the number of questions provided by a particular user. We can pose the following question: Will users provide more answers and fewer questions after this policy change? We exclude this design from our study because the policy change was recent and data collected after the treatment event was not sufficient to perform the analysis.

## 3.5 Conclusions

In this chapter, we discuss QEDs as a set of powerful tools to obtain causal knowledge from observational data and present results from one of their early applications to rich data sets from social media platforms. Specifically, we present three different applications of QEDs to answer three different questions using data from a social media platform, the Stack Overflow website, a peer-to-peer knowledge sharing platform. Our results suggest no significant effect of having a high-quality answer on the number of subsequent answers. Furthermore, specific badges designed to drive engagement seem to work until the badge is received, however, engagement is significantly reduced after. Finally, we show that answer ordering has no effect on the number of votes leveraging a natural experiment design.

Many of these QEDs assume that all common cause variables are observed. However, in practice, observational data sets can easily miss such variables, resulting in biased causal estimates. For handling latent common cause variables, this thesis suggests a novel and distinct approach by using predictive models in machine learning. The next section describes a novel predictive model for inferring values of ordinal latent variables with distant supervision. We illustrate the use of the model for estimating key demographic variables that are often times unobserved on social media platforms.

# CHAPTER 4

# A GENERATIVE MODEL FOR ORDINAL VARIABLES WITH DISTANT SUPERVISION

Modern modeling frameworks in machine learning have been widely used to infer hidden variables almost always with a predictive goal. In this thesis, we explore the utility of such predictive models in adjusting for unobserved variables that are otherwise unavailable.

For example, external data sources, such as first names associated with birth years, last names associated with ethnicity, and zip codes associated with household income, can be useful predictors of age, ethnicity, and income, respectively. This might enable researchers to adjust for such key variables (e.g., age) when they are unobserved, using other correlated signals (e.g., first name) with predictive models.

In this chapter, we develop a novel predictive model for key demographic variables and use it to estimate key demographic information about the population of Twitter users in the United States. We propose to use the generative model framework to incorporate external data sources in the form of aggregate level statistics (i.e., distant supervision)[1]. Specifically, first, we develop and evaluate a generative model that can infer ordinal latent variables with distant supervision. We show the effectiveness of such a model on estimating age from first names. Second, using this proposed model, we perform the largest sample size analysis of US Twitter user population to estimate several useful demographic information. We estimate the age breakdown for the US population of Twitter users over time and show the representation of each age group compared to their corresponding share in the general internet population. We also report on the size of the teenager users on Twitter, which is overlooked in previous reports. We analyze how different age

[1] Much of the content of this chapter is derived from: H. Oktay, A. Fırat, and Z. Ertem. (2014).

groups engage in different topical conversations on Twitter. We also show evidence of assortative mixing in the *follow* relationship among certain age groups on Twitter. Third, we evaluate another generative model for another key demographic variable (i.e., ethnicity), and use it to estimate ethnicity information about the Twitter user population.

## 4.1 Related Work

Many academic studies have used a variety of data sources and methods to estimate demographic information about users of social media and other internet services. We review them in three distinct categories.

The first category of studies used self-reported demographic information either in their online profile or through user surveys. Sharad et al. [34] linked self-reported demographic information of internet users to their internet browsing activity to estimate different usage patterns among different demographic groups. Mislove et al. [61] identified geography, gender, and ethnicity by using self-reported location and name information on Twitter profiles to estimate the demographic properties of US Twitter users. They simply mapped each first name to the most likely gender and each last name to the most likely ethnicity by using US Census data about last name ethnicity distributions[2]. Pew Research [27] reported different demographic attributes of users of social media websites including Twitter through the analysis of data from phone surveys. Methods in this category, though useful, by definition are limited not only to report on attributes that are self-reported in user profiles but also to exclude populations that cannot be reached with the methodology employed by the study (e.g., Pew Research study ignored users who were less than 18 years old by compulsorily exclusion from their phone surveys).

The second category of studies used annotated data sets to develop supervised models. Rao et al. [83] used manually annotated tweets to predict age, gender, and regional origin about Twitter users with stacked-SVM-based classification algorithms built using linguistic and network-

[2] https://www.census.gov/genealogy/www/data/2000surnames/index.html

structure based features. Due to labeling difficulties, they defined two age categories and manually annotated users with their respective categories: below 30 and above 30. Another study by Zamal et al. [110], similarly, trained *SVM* models from manually labeled tweets focusing gender, age, and political affiliation using features based on profiles of individual users as well as their neighbors in the social graph. Pennacchiotti et al. [78] used gradient boosted decision trees to predict ethnicity, political orientation, and gender of Twitter users. They obtained a training data set by manually annotating user profiles with ethnicity and gender by examining user's profile picture; and by extracting political affiliation from the user's profile information. They used linguistic features captured by a topic model analysis over all tweets of users as well as network-structure based features. Finally, Nguyen et al. [66] used a linear regression and a logistic regression model based on text features to predict age, by modeling it continuous and categorical, respectively. They trained their model on manually labeled training instances of Dutch Twitter users to estimate different linguistic usages among different age groups. Methods in this category require correctly labeled data sets about each individual user to train robust models and inherently limited in the size of training data sets by the resources available for annotation.

The third category of models used distant supervision to inform predictive models. O'Connor et al. [67] proposed a generative mixture model to understand demographic and linguistic variations among Twitter users using geotagged tweets. Zip codes of users were cross-referenced with US Census statistics[3] to obtain the ethnicity distribution reported in the corresponding zip code. This allowed them to discover latent communities in Twitter characterized jointly by linguistic and ethnic properties. In another study, though in a non-social media context, Gallagher et al. [32] used data about baby names from the US Social Security Administration[4] as a prior along with image-based features in a generative mixture model to estimate the age of a person from an image-first-name pair. Finally, Chang et al. [22] used US Census data[3] about last names and

[3] https://www.census.gov/genealogy/www/data/2000surnames/index.html

[4] http://www.ssa.gov/oact/babynames/limits.html

their corresponding ethnicity distributions in a generative mixture model to predict ethnicity of Facebook users from their last names.

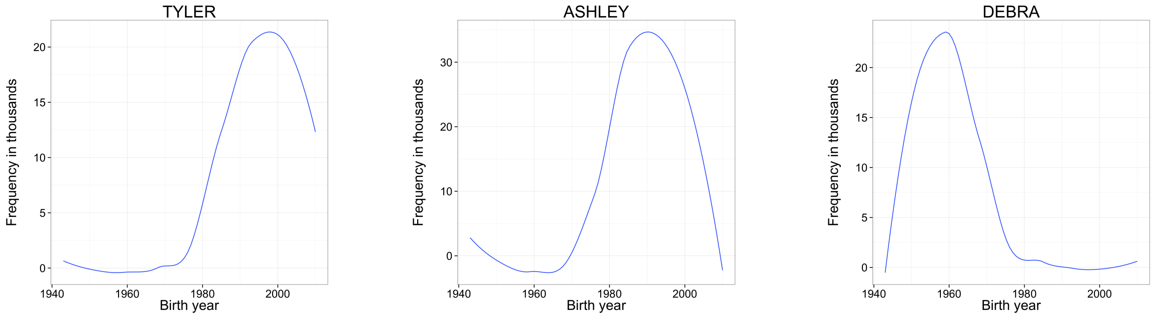
Our proposed method, in this chapter, builds on the work in the third category of models and describes a novel generative model to predict ordinal latent variables with distant supervision such as age and income. In this chapter, we use this model for age estimation using first names. Our distant supervision comes from the Social Security baby name frequency data set. The proposed model removes the need for individually labeled data sets and can estimate variables unreported in user profiles such as age or income.

## 4.2 Age Estimation Using First Name

In this section, we describe a generative model to predict age by considering first names along with Social Security data (SSD) about baby names.[4] The input to the model is a list of first names along with name frequencies from SSD for each birth year. The output of the model is two-fold: (1) an individual-level prediction of age for each name; (2) a population-level prediction of aggregate age breakdown for the entire population.

The U.S. Social Security Administration releases data about the frequency of each baby name for each year. Such data include the frequency of more than 150 thousand different baby names for each year starting from 1881. We show in Figure 4.1 some example first names and how their smoothed frequency values show temporal trends over years. For example, among common names (names that belong to at least 500 thousand people), a person named *Tyler* is less than 30 years old 94% of the time. Similarly, a person named *Debra* is *50-64* years old 78% of the time.

A naïve method to predict age using such a signal would be to independently assign a first name to the corresponding most frequent birth year in the SSD [61]. Then to estimate population-level statistics, we can simply sum up the probability vector of birth years for each name in a given population and then normalize the sum. The suggested naïve analysis assumes that members of the population of interest are independent of each other. Inferring age for each name independently is a less useful indicator of age distribution than inferring collectively when the members of a



(a) Probability that a person named Tyler was born between 1983-2013 is 94%.

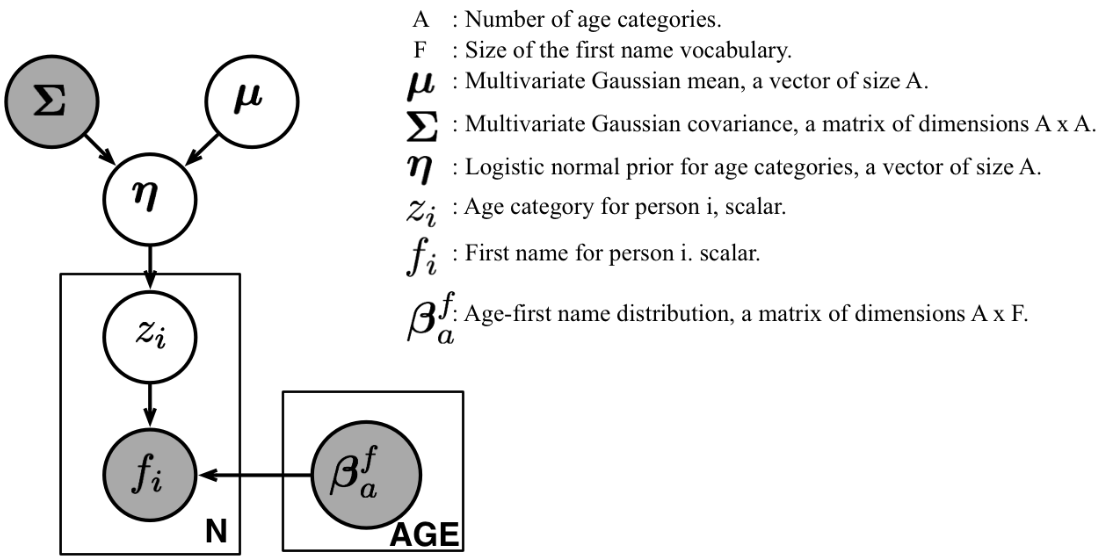
(b) Probability that a person named Ashley was born between 1973-2013 is 88%.

(c) Probability that a person named Debra was born between 1949-1963 is 78%.

Figure 4.1: Smoothed frequencies of example first names over birth years show that baby names in the US show clear temporal trends.

population of interest show assortative mixing [38], and there is an abundance of evidence that social media platforms exhibit high-order assortative mixing (e.g., [17, 60, 65]).

For example, the list of names of students attending a college is clearly not an independent sample, and it is likely that the list of names of college students includes many names that are popular in relatively recent birth years such as Ashley (i.e., high-order assortative mixing). Thus, each name in a population is not necessarily independent from each other. In other words, probabilistically, assume that we have a name in a college student population that is uniformly popular among all age groups. If the inference is performed independently on such a name, the model makes predictions randomly due to the uniform probability. Whereas if the inference is performed collectively by considering other names in the population, the model shifts its prediction towards young age groups. That, in turn, happens to other age groups because many names indicate highly specific age groups (e.g., Debra or Ashley, as in Figure 4.1). The proposed generative model (described in the next section) provides a powerful framework by allowing each name in a given population to inform each other name through collective inference.



$\Sigma$
$\mu$
$\eta$
$z_i$
$f_i$
$\beta_a^f$
N
AGE
A : Number of age categories.
F : Size of the first name vocabulary.
$\mu$ : Multivariate Gaussian mean, a vector of size A.
$\Sigma$ : Multivariate Gaussian covariance, a matrix of dimensions A x A.
$\eta$ : Logistic normal prior for age categories, a vector of size A.
$z_i$ : Age category for person i, scalar.
$f_i$ : First name for person i. scalar.
$\beta_a^f$: Age-first name distribution, a matrix of dimensions A x F.

Figure 4.2: The proposed Bayesian graphical model predicting age given first names. Shaded variables indicate the inputs to the model, and unshaded variables are inferred using the model.

### 4.2.1 Model

We use a generative Bayesian mixture model to estimate the age density of a given population [15, 16, 43, 57, 64]. We model the individual age values as hidden variables, and we model the first names of corresponding individuals as observed variables. Also, parameters corresponding to first name frequencies for each age value are observed (through SSD data). Then, we simply infer the most likely values for hidden variables (individual age values and population mixture) given the observed variables (corresponding first names and corresponding frequencies in each age category).

In our model, we explicitly account for the ordinal relations among the age values by using a logistic-normal prior for age proportions with a predefined covariance matrix [3, 15]. The model enables collective inference [38] through the shared logistic-normal prior among all the names in the population.

More formally, the age prediction model assumes that the first names present in a population of size $N$ arise from the following generative process. Let $A$ be the number of different age categories, and $N$ be the number of users in a given population.

1. For each age value $a \in 1...A$, get a first name frequency distribution from SSD, $\boldsymbol{\beta}^f_a$.
2. Draw $\boldsymbol{\nu}$, from a multivariate Gaussian with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\Sigma}$.
3. Transform $\boldsymbol{\nu}$ with the logistic function to get age proportions, $\boldsymbol{\eta} = F(\boldsymbol{\nu})$.
4. For each person $n \in 1...N$,
    - (a) Draw an age value $z_i$ from $Multinomial(\boldsymbol{\eta})$.
    - (b) Draw the first name of an individual based on age value, $f_i \sim Multinomial(\boldsymbol{\beta}^f_{z_i})$.

The graphical representation of the corresponding model is shown in Figure 4.2. The parameters of the model are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for multivariate Gaussian, and $\boldsymbol{\beta}^f_a$ for multinomial distributions of first names for each age category (i.e., $\boldsymbol{\beta}^f_a, a \in 1...A$, where A is the total number of age categories). We transform multivariate Gaussian values to multinomial parameters using logistic-normal transformation as follows:

$$F(\nu_i) \triangleq \eta_i \triangleq \frac{e^{\nu}_{i}}{\sum_j e^{\nu_j}}.$$

The shaded variables in Figure 4.2 correspond to the observed variables, and unshaded variables correspond to the hidden variables. We set $\boldsymbol{\beta}^f_a$ values from the SSD about baby names.

Age is an ordinal variable where the following inequality holds for an age value $x$: $x - 1 < x < x + 1$. From the posterior probability perspective, such ordinal dependency translates into having similar probability values for age values that are close to each other. For example, as shown in Figure 4.1, if a person's name is *Ashley*, there is a high probability of being born in 1990 and a similarly high probability of being born in 1991 and 1989. By modeling the ordinal dependency among age values, we explicitly model high probabilities for values that are close to 1990 (e.g., 1991, 1989). On the other hand, a person named *Ashley* has a low probability that she was born in 1950, and therefore we explicitly model low probability estimates for values close to 1950 (e.g., 1951, 1949).

We set the parameter for $\boldsymbol{\Sigma}$ as suggested by Agresti [3] to model the ordinal dependency among the age values as $\boldsymbol{\Sigma}_{ij} = \rho^{|i-j|}$, where $0 < \rho < 1$ and $i$ and $j$ are indexes for consecutive possible

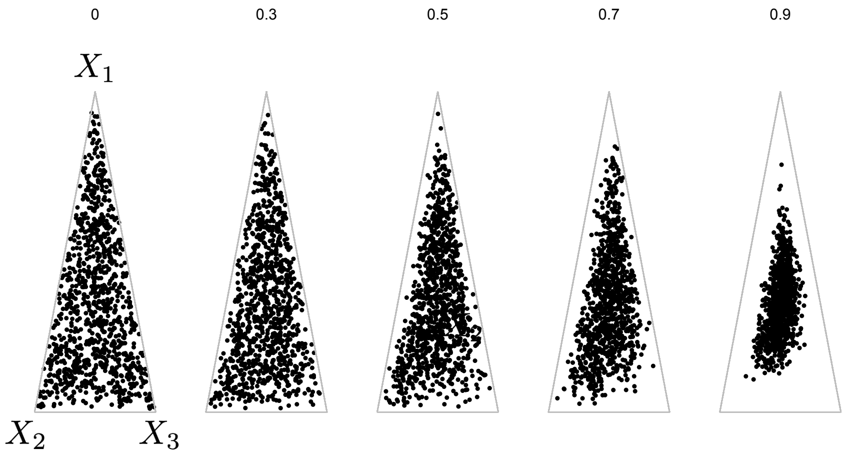


Figure 4.3: Possible values of class mixture proportions for a three-valued variable as the covariance values change for the logistic-normal prior. When $rho$ is small, all values can be possible. As we increase $rho$, the space of possible values is constrained, mimicking the ordinal dependency among values. In the proposed model, this mixture proportion corresponds to the probability vector for the multinomial.

values of the age variable. We show in Figure 4.3 how setting the $\mathbf{\Sigma}$ as suggested by Agresti [3] constrains the possible probability vectors for an ordinal variable with three possible values. Each dot represents a probability vector that sums to 1. In this figure, each corner represents a possible value for the ordinal variable, and the distance from a dot to a corner represents the probability of the corresponding value. The closer the dot is to the corner, the higher the probability for that particular value. We systematically change the $\rho$ value from 0 to 0.9 and plot 1000 possible probability vectors. We observe that as the $\rho$ value for the covariance matrix increases, the values in the probability vector for the random variable become more and more dependent, mimicking the ordinal dependency.

The individual age categories (i.e., $z_i$), the age breakdown of the population (i.e., $\boldsymbol{\eta}$), and the mean values for the logistic normal prior (i.e., $\boldsymbol{\mu}$) are hidden variables and to estimate their values we need to perform posterior inference. Mathematically, we must estimate

$$p(\boldsymbol{\eta}, \boldsymbol{\mu}, z_{1:N} | f_{1:N}, \boldsymbol{\beta}^{f}_{1:A}, \mathbf{\Sigma}) \tag{4.1}$$

Although a logistic-normal distribution can model ordinal aspects of age, parameter estimation and inference are challenging because logistic-normal distribution is not a conjugate prior for the multinomial distribution [15]. Therefore, we use variational methods for parameter estimation by modifying the inference algorithm proposed by Blei et al. [15] for correlated topic models. We report results using the expected value of each hidden variable under the variational inference.

For a given list of names in a population, we infer the age breakdown of the population using $p(\boldsymbol{\eta}|f_{1:N}, \boldsymbol{\beta}^f_{1:A}, \boldsymbol{\Sigma})$. For convenience, we define $\boldsymbol{\pi}_i$ as the estimated age breakdown of user $i$

$$\boldsymbol{\pi}_i \triangleq p(z_i|f_{1:N}, \boldsymbol{\beta}^f_{1:A}, \boldsymbol{\Sigma}) \tag{4.2}$$

where $\boldsymbol{\pi}_i$ is a vector of size $A$ whose $a^{\text{th}}$ element expresses the probability that the user is in age group $a$. This representation supports custom age categorization by simply defining new age category boundaries and summing up the corresponding values in $\boldsymbol{\pi}_i$, for each category.

In Section 4.2.4, we report results on the relationships between age groups. To achieve this, we define the following matrix between a pair of users, ($n_1$, $n_2$), in a particular relationship,

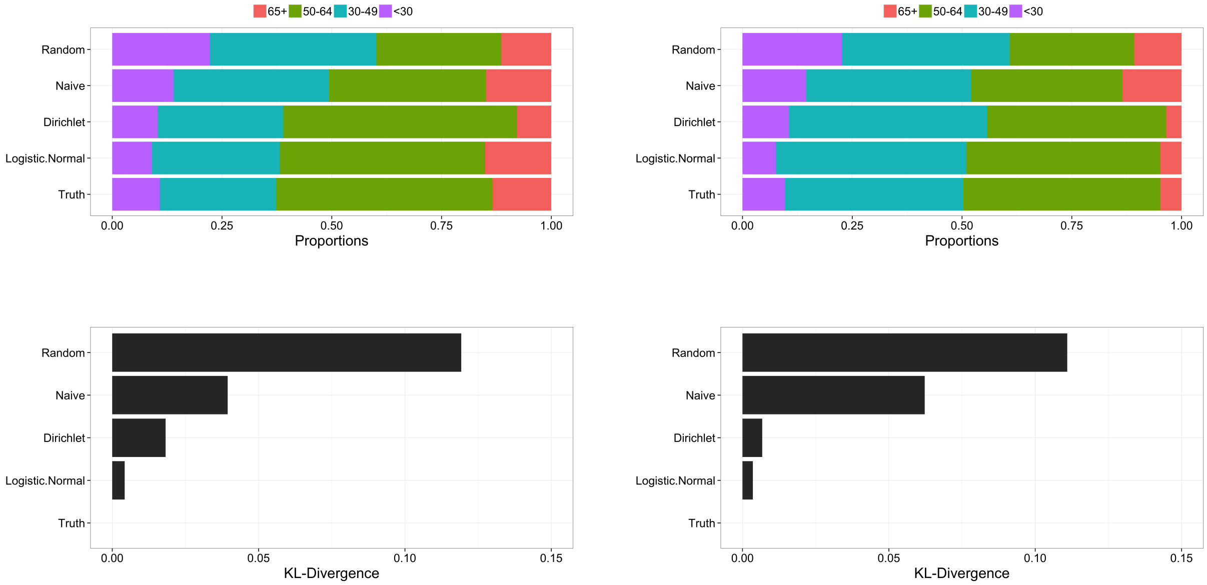
$$\boldsymbol{M}_{n_1,n_2} \triangleq \boldsymbol{\pi}_{n_1}\boldsymbol{\pi}^T_{n_2}, \tag{4.3}$$

where $\boldsymbol{M}_{n_1,n_2}$ is an $AXA$ matrix whose $(a_1, a_2)^{\text{th}}$ entry expresses the probability that user $n_1$ is in age group $a_1$ and user $n_2$ is in age group $a_2$.

### 4.2.2 Limitations and assumptions of the model

First of all, by using the data compiled by the US Census, we assume that the populations we use in this model to make inference are sub-populations included in the US Census. For example, using this model on populations from the UK might fail.

Second, the generative process assumes that the first names are independent of the population mixtures given birth years. This suggests that we assume being in a certain population only depends on birth year, not first name. For example, if people whose first names start with *T* are more likely to be in a certain population this assumption would be violated.



(a) Results on an example data set

(b) Results on another example data set

Figure 4.4: Two different examples from our experimental evaluation in which the performance of the logistic model roughly corresponds to the average performance among 1000 test runs. A logistic-normal model estimates the age distribution of populations better than the natural alternative models. Each column represents results from an example data set. The figure on the top in each column shows the actual predictions of each model. The figure at the bottom shows the KL-divergence of predicted mixture distributions to the true mixture.

Third, the generative process assumes that, given a population mixture, birth year values of any two first names are independent of each other. For example, if a population includes grandchildren only when their grand parents are also in the population, this assumption would be violated.

### 4.2.3 Experimental Evaluation

We evaluate the proposed model by using voter registration data. We gather 154,016 voter registration records from Ohio in which we have the ground truth data with first names and corresponding birth years. We compare performance of the proposed method to the performance of these following natural alternative models in estimating the population level statistics:

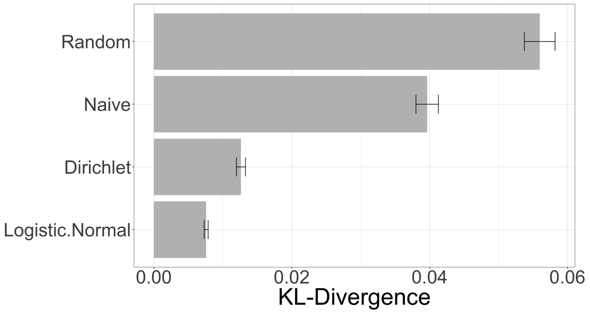


Figure 4.5: The model with logistic-normal prior achieves the smallest (i.e., best) KL-divergence as it both models age as an ordinal variable and performs collective inference.

**Truth** is the ground truth data from voter registration.

**Random** is randomly assigning a person to a birth year, and then calculating the aggregate proportions.

**Naïve model** is a model that simply aggregates the probabilities in the SSD. (This corresponds to the independent inference of names, as described in the text.)

**Dirichlet prior model with smoothing** is a mixture model with collective inference, but models age as a categorical variable (i.e., not ordinal). We smooth the predictions of this model to make it more suitable for estimating an ordinal variable. (We note that the performance without smoothing is much worse than with smoothing.)

**Logistic-normal prior model** is the proposed model that explicitly accounts for the ordinal dependency among age values.

We evaluate the proposed model on estimating the population-level age proportions. We systematically sample populations with different mixtures of birth years using the ground truth data. We create 1000 different data sets, each of which includes 10,000 first names, and we use the proposed model to estimate population-level statistics. We evaluate performance at an age category level, where we combine birth years into categories by summing the density values of corresponding birth years. For example, we estimate the density for the *18-29* age group by summing all the birth year values corresponding to that group. We define four age categories: *<30, 30-49, 50-64,* *and 65+*. For the evaluation metric at the age category level, we use KL-divergence (the smaller the value, the better the performance).

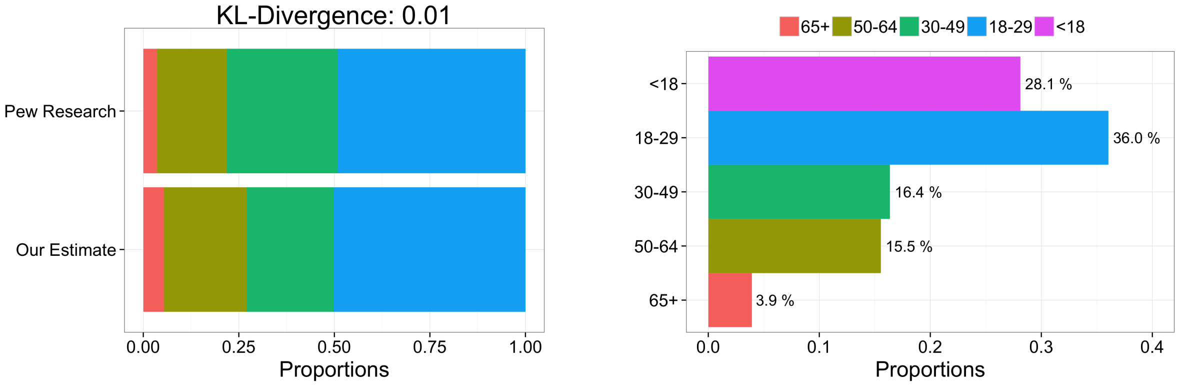
We expect that models utilizing collective inference should outperform naïve models, especially when the population of interest is skewed from the population that is representative of SSD. In Figure 4.4, we show two test runs with different age distributions. In the top row; we show actual predictions, and in the second row we show the KL-divergence between estimations and true distribution. In the top figures, the bottom bar corresponds to the true age distribution, and the rest above are predictions from different models (random, naïve, smoothed Dirichlet, and logistic-normal, respectively).

In Figure 4.5, we plot the average KL-divergence of each model's predictions for 1,000 test cases. We observe that models that utilize collective inference (i.e., logistic-normal and Dirichlet) perform better than models that do not utilize collective inference (i.e., naïve and random). We also report that the logistic-normal prior, which models age as an ordinal variable, performs better than Dirichlet prior, which models age as a categorical variable. These experimental results suggest that both collectively inferring posterior probability and explicitly modeling the ordinal dependency among age values improve the estimation of population-level statistics.

### 4.2.4 Application to Twitter

In this section, we apply the proposed model described in Section 4.2.1 to US Twitter users. First, we replicate a popular Pew Research study [27] about Twitter demographics, and eliminate a methodological limitation in their study. Second, we report on how the age diversity of US Twitter users has changed over time.

With the model described in Section 4.2.1, we can accurately estimate the relative age breakdown of Twitter users over time. First, we calculate the age category breakdown of a Twitter user base snapshot from December 2012 and compare our findings to estimates in a Pew Research Report about US Twitter demographics [27]. We define our categories as follows: *<18, 18-29, 30-49, 50-65, and 65+*, as suggested in the Pew Research report. We note that first name is an optional



(a) Pew Research comparison

(b) Age breakdown for full data

Figure 4.6: Estimates of the proposed method are comparable to the findings in the Pew Research report. Other than their report, we also provide estimates for the age group *<18* that such category might be the second largest age group among US Twitter users. The Pew Research overlooked estimates of this category due to methodological limitations.

feed on a Twitter profile. We assume that Twitter users who report their name in their profile report their correct name, and who do not report their name are missing at random.

In Figure 4.6a, we compare our findings based on a random sample of one million random Twitter users in the US to the findings of Pew Research and show that we are able to closely replicate their conclusions. Similar to their conclusions, according to our analysis the largest age group of the US Twitter users is in the *18-29* age group and the least active age group is *65+*. However, due to a methodological limitation, the Pew Research study only focused on Twitter users over 18, completely excluding users below 18 years old (i.e., they could only perform phone surveys for users above *18*). Therefore, in Figure 4.6a, we compare our findings for the age categories over *18* years old.

In contrast to the Pew Research methodology, our proposed methodology does not have a limitation and can give estimates for any age category, including *<18*. Figure 4.6b shows our estimate for all age categories, and we find evidence that US Twitter users who are less than *18* years old might be the second largest age-category among US Twitter users.

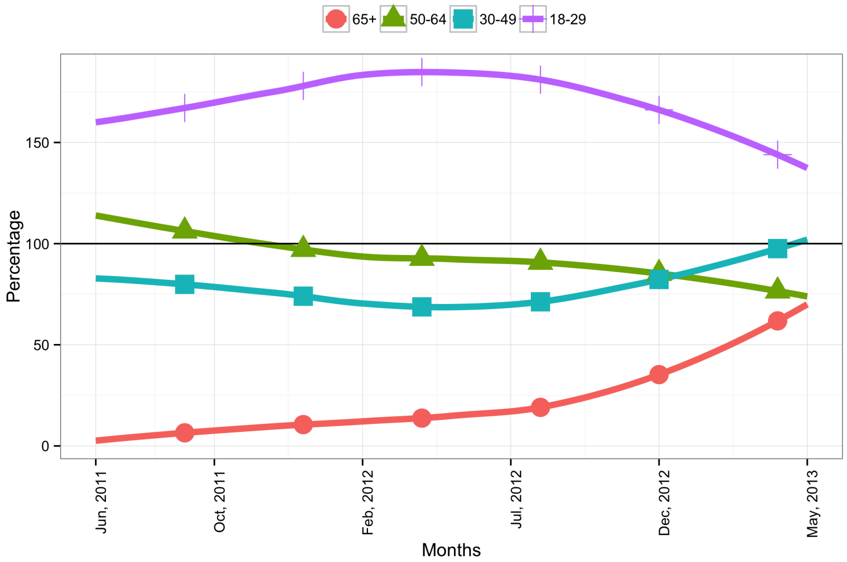


Figure 4.7: We compare different age groups in the US Twitter users to the general US internet population as reported by the Pew Research. We observe that the largest age group on Twitter has been *18-29*, and members of this age group have consistently been over represented compared to their population in the general US internet population as reported by Pew Research [27].

**US Twitter user population age breakdown:** Also, with the proposed methodology, we can estimate how different age groups are represented on Twitter over time. We analyze a random sample of US Twitter users of size 100K for each month starting from June 2011 to May 2013. In Figure 4.7, we compare the saturation of each age group over *18* years old to the age breakdown of internet users obtained from the Pew Research report [27]. We observe that among US Twitter users, age group *30-49* is almost saturated (i.e., has its fair share of users compared to the US Internet population). Age group *18-29* is over-represented, and other age groups are under-represented—though *65+* seems to be increasing rapidly, in the first half of 2013.

**Activity of different age groups throughout the day:** With the proposed method, we can also analyze which age group is actively using Twitter at what time during the day. We randomly sample 170K+ tweets posted at different hours from a one-month interval. In Figure 4.8, we plot the daily activity of different age categories. We observe that US Twitter users under *18* and above *65* might be active on Twitter at completely different times of the day. Twitter users above 65

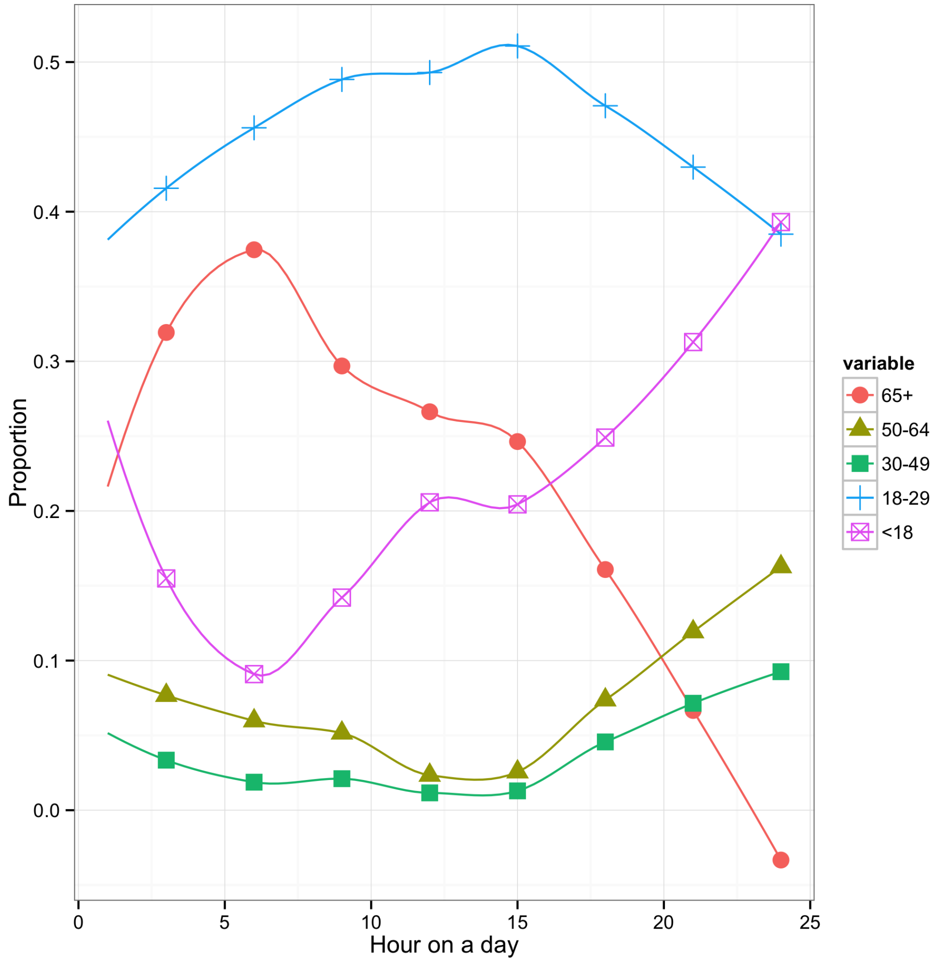


Figure 4.8: During the day different age groups post tweets at different times in the US. Users over *65* and users under *18* use Twitter at completely different times.

seem to be most active early in the morning. In contrast, Twitter users under *18* seem to be posting Tweets later in the day. For the largest age group on Twitter, age category *18-29*, Twitter usage peaks later in the afternoon. The age groups *30-49* and *50-64* show a similar pattern where their usage peaks around late midnight.

**Follower Analysis:** We also perform some example case studies on US Twitter users to show various potential applications of the model. First, we focus on age breakdown of followers of arguably popular US Twitter users, and we compare such breakdown figures with respect to the breakdown of general US Twitter users.

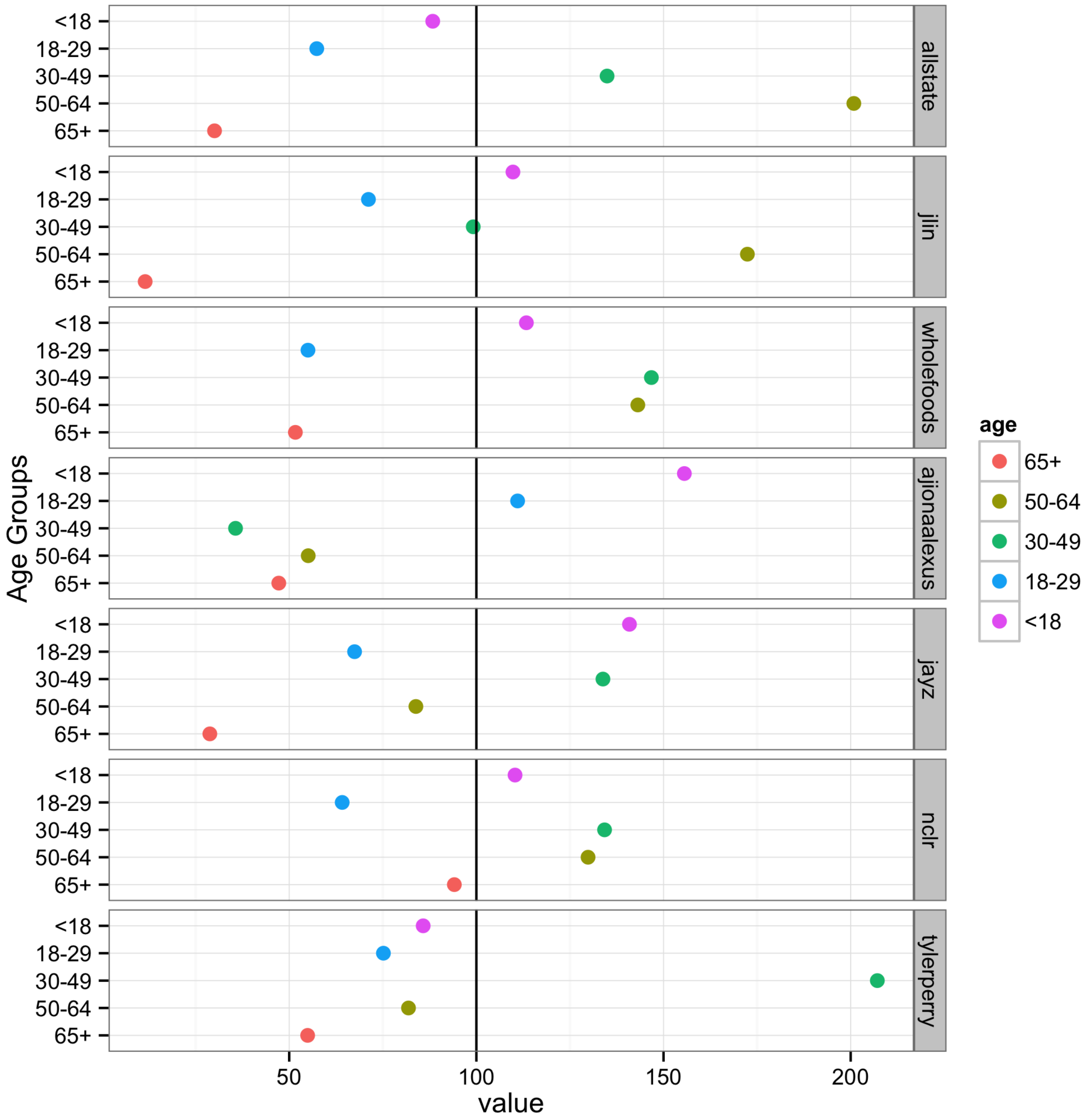


Figure 4.9: We perform a case study on US Twitter users focusing on followers of a set of popular Twitter users. Different Twitter users have followers from different age groups.

Figure 4.9 compares the age breakdown for each group of followers to the age breakdown of a random sample of US Twitter users. If an age group is well-represented with respect to the general Twitter population, the dot should be on the solid vertical lines, which intersect with 100%. Dots to the left of this line indicate under-representation, and dots to the right side of this line indicate over-representation of the corresponding age category with respect to general Twitter population.

For example, we observe that the *50-64* age group is more than four times more represented in the followers of *Allstate*, an insurance company, than its corresponding share in the general US Twitter population. We see a similar trend for followers of *Whole Foods*, a national chain of grocery stores in the US, and also for followers of *NCLR*, an NGO organization for Hispanic rights located in Chicago.

On the other hand, for followers of *Ajiona Alexus*, a teenage celebrity, we observe that Twitter users under age *18* are over-represented compared to their corresponding proportion on Twitter. Different Twitter users attract followers from different demographic groups and this proposed model can estimate such population-level breakdowns for the followers of different users.

**Topical Analysis:** Using the proposed model, we can also analyze the breakdown for different age groups engaging in conversation around specific topics on Twitter. In Figure 4.10, we compare the age breakdown of users given a specific topic to the general age breakdown of US Twitter users. As explained above, 100% indicates perfect representation of each age group, smaller values indicate under-representation, and higher values indicate over-representation.[5]

We observe that the age groups are represented differently in different conversations. For example, for the *immigration act* conversation on Twitter, measured by the relevant set of hashtags, we observe that Twitter users of age *65+*, *<18*, and *30-49* are over-represented (i.e., provided more posts than their expected share by their general proportion on Twitter). We also observe that users of age *18-29* are under-represented in the immigration act conversation on Twitter.

Users in age group *50-64* seem to be actively participating in conversations about *fiscal cliff, gun regulations, the blizzard nemo, boston-strong campaign* and *nba-finals*. We also observe that users under *18* are also participating in conversations about *nba finals*.

[5] We gather data for each of these topics using the *Twitter Firehose API*. We filter tweets containing relevant keywords and hashtags. See Section A.1 for a detailed list of hashtags used for each topical analysis.

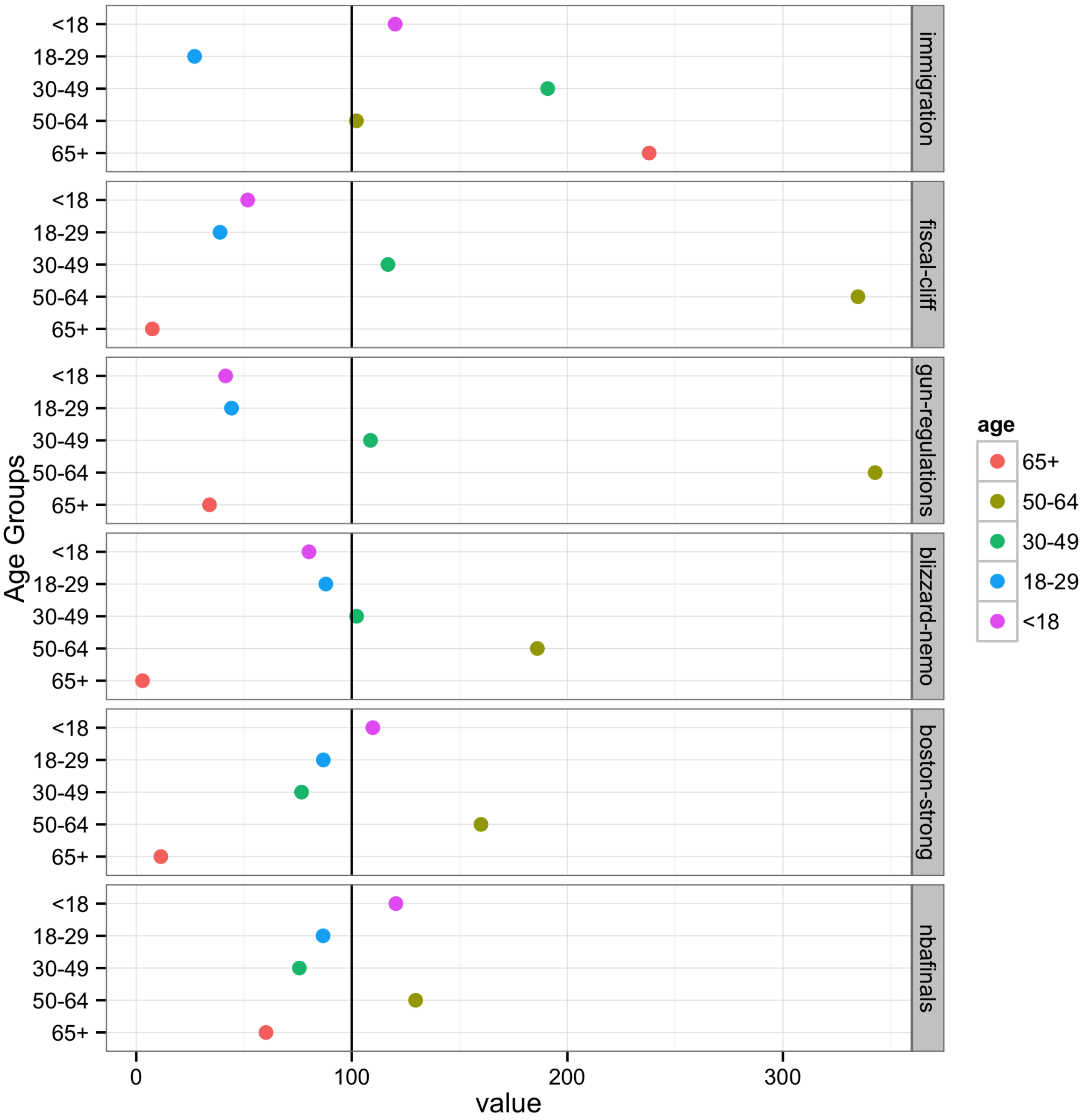


Figure 4.10: We perform a case study on US Twitter users engaging in specific conversation. Different topics attract users from different age groups as participants.

**Follow relationship analysis:** We analyze how different age groups interact with each other with respect to the follow relationship on Twitter. We sample 100,000 random follow relationships from the Twitter social graph. We then use our model to predict the age of each user in the sample. For each relationship, we calculate an interaction matrix by simply taking the cross-product of

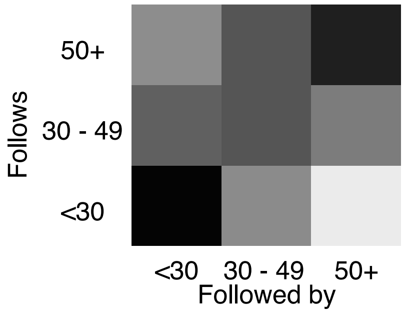


Figure 4.11: Following relationship between different age groups. Darker cells indicate that age group pairs follow each other (i.e., y follows x) more frequently than expected by chance. We observe that *<30* and *50+* age groups show assortative mixing in their relationship, however, *30-49* age group shows diverse assortative mixing.

age prediction vectors corresponding to users involving in a follow relationship, as described in Equation 4.3. We calculate the interaction matrix for the population by aggregating individual matrices, as follows. Let $\kappa$ denote the set of pairs in a relationship,

$$\hat{\boldsymbol{M}} \triangleq \frac{1}{|\kappa|} \sum_{(n_1,n_2)\in\kappa} \boldsymbol{M}_{n_1,n_2} \tag{4.4}$$

where $\hat{\boldsymbol{M}}$ is an $AXA$ matrix whose entries denote the estimated proportion of relationship being of a particular pair of age groups.

We calculate an average interaction matrix by aggregating the matrices for random edges sampled from the Twitter graph, using Equation 4.4. We normalize the average interaction matrix, with the expected interaction matrix by chance, $\boldsymbol{M}^* \triangleq \boldsymbol{\eta}^*\boldsymbol{\eta}^{*T}$, where $\boldsymbol{\eta}^*$ is the population age breakdown. This practically corresponds to sampling random pairs from the population.

Figure 4.11 shows the normalized interaction matrix as a heat map. In this matrix, the age group in the y-axis follows the age group on the x-axis. The darker the cell, the higher the interaction between the corresponding pair of age groups. We observe that the age groups *<30* and *50+* show high assortative mixing (i.e., users tend to follow mostly users within their own age group), whereas the age group *30-49* shows less assortative mixing.

| **Name** | **Rank** | **Count** | **% in group** |
|---|---|---|---|
| African American | | | |
| Washington | 138 | 163036 | 89.7% |
| Jefferson | 594 | 51361 | 75.2% |
| Booker | 902 | 35101 | 65.6% |
| Asian/Pacific Islander | | | |
| Zhang | 963 | 33202 | 98.2% |
| Huang | 697 | 44715 | 96.8% |
| Choi | 872 | 57786 | 96.4% |
| Hispanic | | | |
| Barajas | 989 | 32147 | 96.0% |
| Orozco | 690 | 45289 | 95.1% |
| Zavala | 938 | 34068 | 95.1% |
| White | | | |
| Yoder | 707 | 44245 | 98.1% |
| Krueger | 863 | 36694 | 97.1% |
| Mueller | 467 | 64305 | 97.0% |

Table 4.1: Frequently occurring last names for different ethnic groups.

## 4.3 Ethnicity Estimation Using Last Names

Ethnicity is another key demographic variable that is unobserved in many social media platforms. In this section, we present another generative model to predict ethnicity by considering the last name along with US Census Bureau statistics about last name ethnicity frequencies. The input to the model is a list of last names and data from the US Census Bureau statistics (i.e., the frequency of last names in each ethnic group), and output of the model is an ethnicity prediction for each last name, and ethnicity mixture for the corresponding population.

The US Census Bureau publishes data about frequently occurring last names under its Genealogy Project[6]. The data are compiled from almost 270 million responses to the 2000 Census. This data includes last names, their corresponding total count in the population, and their individual ethnic breakdown (i.e., proportions for each ethnicity).

Last names can provide a strong signal for inferring ethnicity. For example, having the last name *Washington* is a strong signal of Black ethnicity (more than 89% of people with last name

[6] http://www.census.gov/genealogy/www/data/2000surnames/index.html

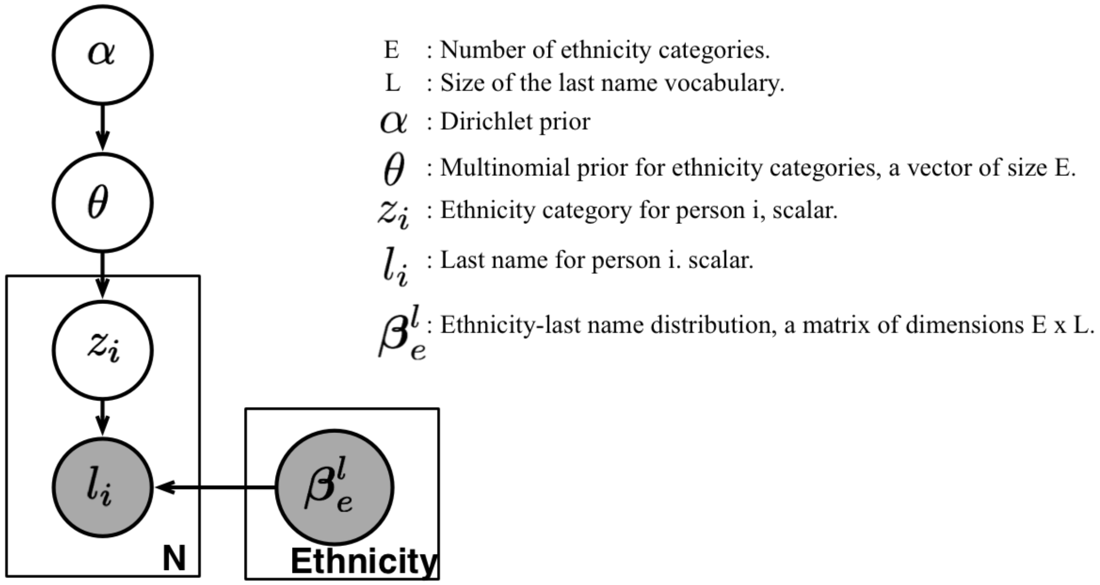

E : Number of ethnicity categories.
L : Size of the last name vocabulary.
$\alpha$ : Dirichlet prior
$\theta$ : Multinomial prior for ethnicity categories, a vector of size E.
$z_i$ : Ethnicity category for person i, scalar.
$l_i$ : Last name for person i. scalar.
$\boldsymbol{\beta}_e^l$ : Ethnicity-last name distribution, a matrix of dimensions E x L.
$\alpha$
$\theta$
$z_i$
$l_i$
$\boldsymbol{\beta}_e^l$
N
Ethnicity

Figure 4.12: Probabilistic Bayesian model to estimate ethnicity proportions.

*Washington* are of Black ethnicity, as shown in Table 4.1). Having the last name Zhang is a strong signal of Asian ethnicity. Having the last name *Krueger* is a strong signal of White ethnicity.

Similar to the age-prediction model proposed earlier, Chang et al. [22] proposed a probabilistic Bayesian model to predict the ethnicity information using last names. Such a model collectively infers the ethnicity of a given population by allowing inference about each name in a given population to be informed by inference about every other name. For example, the last name *Mack* is almost equally likely to be of White ethnicity or of Black. However, if this last name is in a population where there are many other people named *Yoder* or *Krueger* (which are last names that are more likely to belong to people of White ethnicity than of other ethnic groups), then *Mack* is more likely to be of White ethnicity than of Black ethnicity. Instead, if this name is in a population where there are many people named *Washington* or *Jefferson*, then *Mack* is more likely to be of Black ethnicity.

The collective inference is performed on a generative mixture model, as shown in Figure 4.12. Instead of learning the $\boldsymbol{\beta}_e^l$ parameters (distribution of last names given ethnicity), the values from the US Census Bureau statistics are used in the model. As shown in the graphical model in Fig-

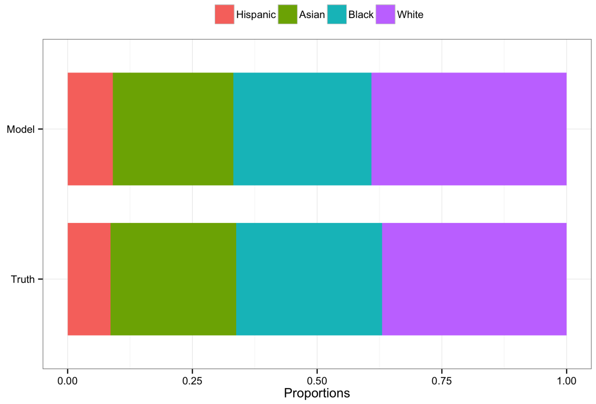


Figure 4.13: Ethnicity breakdown of a data set from Freebase.com. A Bayesian model accurately estimates the proportions.

ure 4.12, last names and the beta ($\beta_e^l$) parameter (i.e., last name frequencies from the US Bureau Statistics), are observed, as indicated by shaded nodes, and theta $\theta$ (i.e., ethnicity distribution of a given population), $\alpha$, and $z_i$'s are inferred through collapsed-Gibbs sampling. We focus on the main four ethnic groups that cover most of the US population: *White, Black, Asian/Pacific Islander* and *Hispanic*[7].

### 4.3.1 Experimental Evaluation

To evaluate this probabilistic Bayesian model, we gather names and corresponding ethnicity information from Freebase.com[8], which is a large knowledge base for structured data compiled from various sources. We focus on Americans of White, Black, Asian/Pacific, or Hispanic origin and put together a data set that contains 3,536 people from the USA. Among these users, 3,076 of them have last names that are present in the US Census Bureau statistics.

First, we test our model on the whole Freebase data set. Figure 4.13 compares the model estimate to the actual ground truth. We use KL-divergence to measure the distance between the

[7] We use the same ethnic categories as the ones in the US Census Bureau data merely because of the availability, though the methodology discussed can be used with any categorization.

[8] http://www.freebase.com

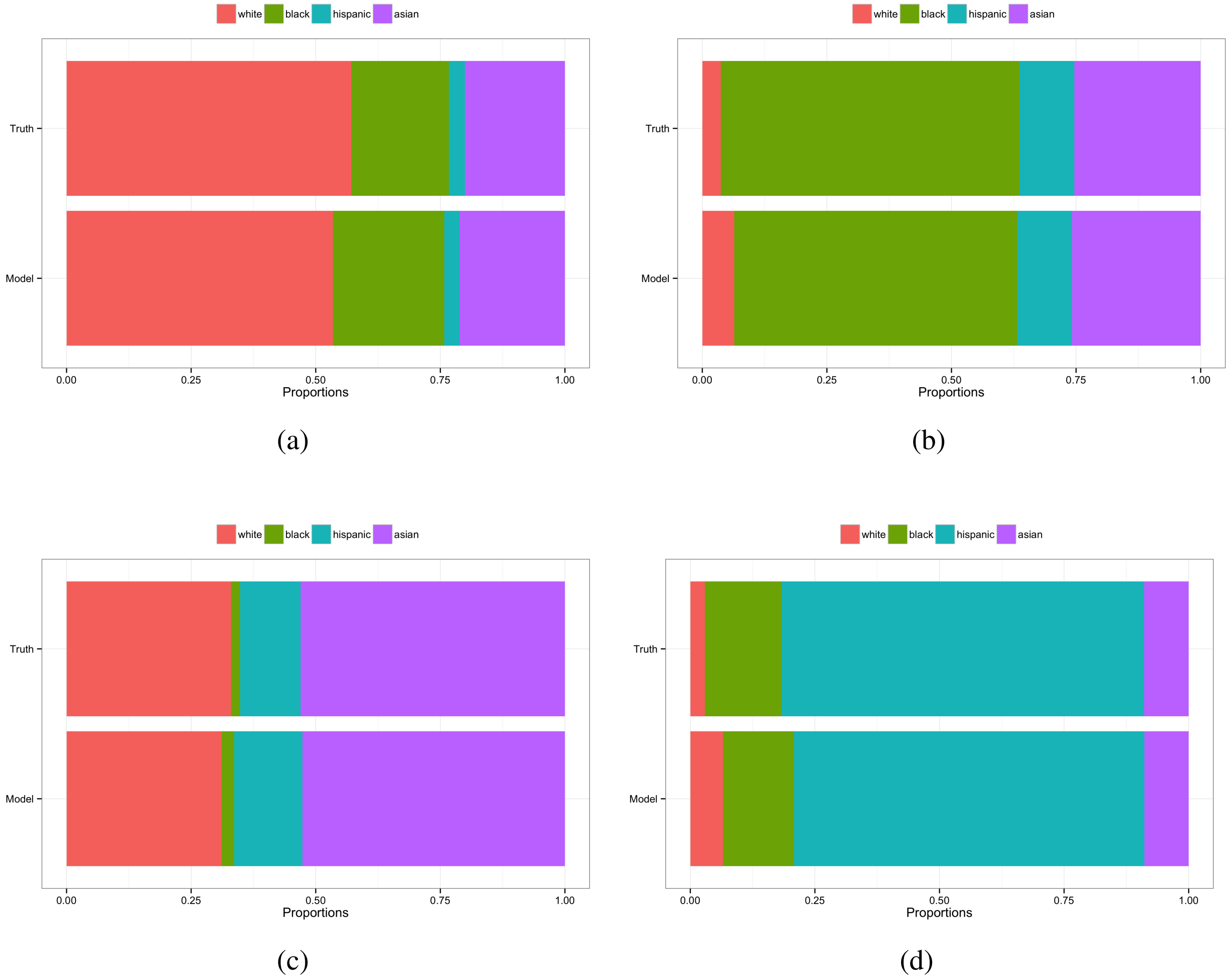
estimated probability distribution and the true distribution. The KL-divergence between the ground truth and the model estimate is less than 0.001. The model is accurately able to estimate the ethnicity proportions for the overall Freebase population.

We further evaluate the Bayesian model for cases where the training distribution (i.e., ethnicity distribution implied by the US Census Bureau statistics) and the test distribution (i.e., ethnicity distribution of a given population) are considerably different from each other. To generate test cases with a skewed distribution, we first randomly sample a mixture proportion from a Dirichlet distribution. Then, we sub-sample from the Freebase data set to satisfy this sample mixture proportion. By randomly changing the values of the Dirichlet distribution, we generate data where ethnicity values are skewed (i.e., people of one particular ethnicity dominating the whole population). We run this experiment on 1000 different list of names, with different ethnicity proportions. We use collapsed-Gibbs sampling for the inference that runs 1,000 iterations.

The average KL-divergence between the true distribution and the estimated distribution is 0.02. Figure 4.14 shows the results of four example data sets where the underlying population is skewed towards one specific ethnicity. We provide evidence that the model is quite accurate even when there is a considerable skew in ethnicity distributions.

### 4.3.2 Application to Twitter

We apply the Bayesian model for ethnicity to US Twitter user data. We present how diversity on Twitter changes over time, and how saturated each ethnicity in US Twitter user base compared to addressable internet population. Diversity in terms of ethnicity on Twitter changes over time as new users join to the platform. The Bayesian model allows us to get predictions for each time snapshot, and to temporally estimate the ethnicity distribution. We note that the last name is an optional feed on the Twitter user profile, and we assume the information is correct when present and missing at random when not. Starting from June 2011, for each month we randomly sample three thousand users who listed their location as the US in their Twitter profile.



(a) (b)

(c) (d)

Figure 4.14: Four example sub-samples of Freebase data with skewed ethnicity distributions compared to the corresponding true density. The Bayesian model accurately estimates ethnicity even though the test distribution is skewed.

Figure 4.15 shows the ethnic proportions on Twitter over time divided by the corresponding proportion of that ethnicity in addressable internet population. We use the estimated ethnic breakdown of households with an internet connection from the National Telecommunications and Information Administration report on the Networked Nation[9]. We observe that Black and Hispanic US Twitter users are over-represented, whereas White and Asian users are under-represented in the

[9] http://www.ntia.doc.gov/report/2008/networked-nation-broadband-america-2007

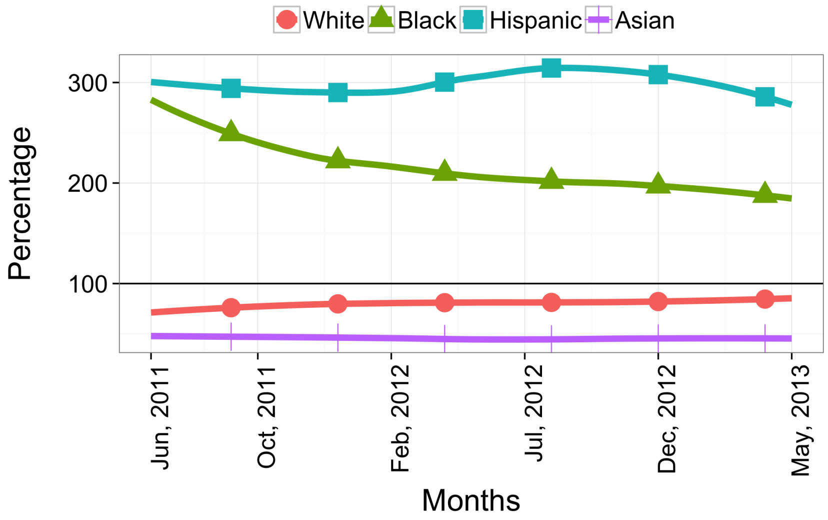


Figure 4.15: Comparison of ethnic proportions of US Twitter users to the addressable internet population shows that Black and Hispanic users have been over-represented, whereas White and Asian users have been under-represented on the platform.

US Twitter population. These results appear to be in agreement with many other news articles and blog posts indicating the over-represented minority population among US Twitter users (especially among early adopters of Twitter).[10]

## 4.4 Conclusions

In this chapter, we develop a novel predictive model based on generative models to predict ordinal variables with distant supervision. We use this model to predict age using first names in a given population. Our model is distantly supervised in the sense that it uses Social Security baby name data for each year as a prior. This model provides accurate population level predictions as well as individual-level predictions using a generative model framework with collective inference. We show the benefit of explicitly modeling the ordinal dependency among age groups.

[10] http://www.businessinsider.com/twitter-study-results-2010-4

We apply our model to estimate demographic information about US Twitter user age groups. First, we closely replicate a Pew Research report about the active age groups on Twitter, finding that *18-29* age group is the most active. Limited by the user groups they could reach by surveying people on the phone, Pew Research study reports age groups older than *18* (i.e., they cannot legally survey users younger than *18*). Our method is free from such a limitation and we find that less than *18*-year-old is the second most active age group on Twitter.

Our results also indicate that users older than *50* are under-represented; conversely, users in the *18-29* age group are over-represented compared to their respective presence in the general internet users population. We identify times in a day each age group is active on Twitter, finding that users in *65+* age group and *18-29* age group are active on Twitter at completely opposite times during the day. We show evidence that different age groups engage in different topics and follow different users. Finally, we find that young and senior Twitter users exhibit high-order assortative mixing in their follow relationships, whereas middle age users follow users from diverse age groups.

We develop a simple yet effective predictive model for ordinal variables with distant supervision. We apply this model to estimating a key demographic variable (i.e., age estimation using first names). We also evaluate another related model for estimating categorical variables (i.e., ethnicity estimation from last names). We argue that such predictive models with distant supervision can estimate key demographic variables. We propose that for causal questions where age or ethnicity might be unobserved, estimations from predictive models, such as presented in this thesis, can help adjust for their effects. Since every estimation process inherently includes error, we point that such estimations would be noisy and we characterize potential ways to deal with such noise. In the next chapter, we propose to use predictions from these models to improve causal estimation by explicitly handling unobserved variables.

# CHAPTER 5

# USING LATENT VARIABLE MODELS TO ADJUST FOR UNOBSERVED CONFOUNDING VARIABLES

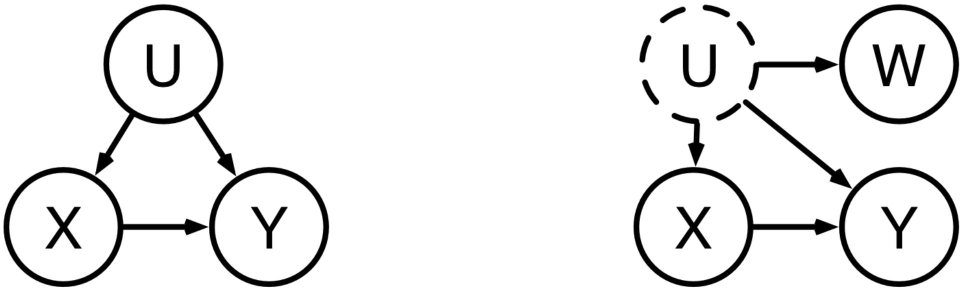
Confounding variables can bias estimates of treatment effects obtained from observational data. For example, when estimating the effect of activity level on weight gain, age is a potential confounding variable because it might affect both activity level and weight gain. Ignoring confounding variables introduces bias in estimates of treatment effect [74, 97], and conditioning on confounding variables is a common approach to adjust for this bias.[1]

However, almost all theoretical principles and practical methods that adjust for confounding variables make the assumption that such variables are measured without error. This conflicts with a harsh reality of empirical analysis: nearly every variable is measured with some degree of error due to misspecification, instrumentation, or recording [92]. When a confounding variable is measured with error, mere conditioning might be unable to completely remove confounding bias [23, 71].

Kuroki et al. [47] proposed algebraic and graphical methods to adjust for measurement error in observed confounding variables. Their proposed method (i.e., effect restoration) assumes knowledge of which variables are confounders, values for their proxy variables, and the corresponding error distributions between proxy and confounding variables.

For example, in the graphical model structure in Figure 5.1b, $X$ is a treatment, $Y$ is an outcome, and $U$ is an unobserved confounding variable. $W$ is a version of $U$ measured with error. While estimating the joint distribution $P(X, Y)$, the proposed method re-assigns the effect of $W$ on $X$ and $Y$ by using the error distribution. When the specific underlying graphical structure in Figure 5.1b

[1] Much of the content of this chapter is derived from: H. Oktay, and D. Jensen (2017).



(a) Confounding variable

(b) Confounding variable with measurement error

Figure 5.1: Graphical model representation of a domain where U is a common cause for X and Y. Figure 5.1a represents perfect observation of a confounding variable. Figure 5.1b represents observing the confounding variable with measurement error.

is assumed, such re-assignment can adjust for the measurement bias. This underlying structure assumes the prior knowledge of confounding variables.

However, in practice, it is far from straightforward to conclude that the assumed structure holds given a treatment, outcome, and a proxy variable. Furthermore, it is unclear whether the adjustment made by effect restoration still provides a bias-free estimate when the underlying generative process does not correspond to the structure in Figure 5.1b, which Kuroki et al. [47] consider out-of-the scope in their study.

In this chapter, we characterize the behavior and extend the use cases of effect restoration for given (X, Y, W) triplets. First, we characterize bias and variance trade-off for effect restoration for all possible underlying structures for X, Y, and W. Second, we formulate the sufficient conditions using d-separation rules to identify the underlying structures of X, Y, and W. Third, we present evidence that effect restoration can reduce bias in real-world experiments. Finally, we show evidence that effect restoration can be used with predictive models to reduce bias by adjusting for unobserved confounding variables. This idea assumes that we have a priori hypothesis about confounding variables and for each such variable we can develop a latent variable model using other observables to estimate its values.

## 5.1 Related Work

Causal sufficiency, a common assumption in casual discovery, states that the set of measured variables include all of the confounding variables for all pairs of variables in a given domain [74, 97]. Many methods focusing on identifying the treatment effect of a particular treatment-outcome pair require a particularly expansive version of causal sufficiency to guarantee a consistent causal estimate: all potential confounder variables of a treatment and outcome pair are *perfectly* observed (i.e., observed with no error) [35, 80, 100].

However, Scheines et al. [92] recently pointed out that in almost any empirical data collection setting, measurement error accounts for some variation in variables due to errors in instruments or recordings. By employing a simulation study, the authors showed that a standard causal discovery algorithm suffers especially in the edge orientation stage from measurement error. In a simpler setting, to estimate one causal effect with measurement error Kuroki et al. [47] proposed theoretical principles of an effect restoration method by extending the *do*-calculus for a specific underlying structure. We employ simulation analysis to enhance this theoretical work with empirical evaluations. We also relax the specific structure assumption and characterize the bias-variance trade off for the adjustment method under other simple graphical structures.

Although not explicitly in causal estimation context, *measurement models* for covariates in regression models have been extensively studied by statistics community, often referred as errors-in-variables [21, 23]. The work presented here differs from models for errors-in-variables because we use graphical models and because we explicitly focus on interventional distributions by using the *do*-calculus, (i.e., $p(y|do(x))$). Effect restoration with graphical models can be mapped into regression analysis only when certain causal assumptions are made [47]. The benefit of explicit causal consideration is to identify measurement models that are compatible with transportability [76], transferring information learned from analysis in one environment to another environment. Here, by building on the explicit causal work by Kuroki et al. [47], and by explicit use of graphical models, we focus on the effects and specification of measurement models for causal estimation.

Furthermore, we explore the use of effect restoration to account for bias from unobserved confounder variables when used with predictive models. Two main methods often used in the literature to adjust for unobserved confounders are instrumental variables where external variables such as weather conditions serve as randomization for the treatment [4] and with-in subject designs, where an experimental subject serves both as a control and a treatment subject for herself [93, 85, 100]. In this chapter, for the same goal, we propose a different approach. We suggest accounting for unobserved confounding variables by using their predictions from independent predictive processes along with their corresponding error distributions in such processes.

Our proposed approach might resemble transfer learning approaches in machine learning where a target learning task is performed by using knowledge obtained from previously related tasks [53, 70, 108]. Early work in this field defines a meta-learning task to capture informative priors through joint inference, that then can inform a new classification task [26, 82]. Recent work includes deep representation learning so that higher-level learned features are transferable across many classification tasks [12, 109]. The main focus in the related literature about transfer learning appears to be virtually classification and prediction tasks as oppose to effect-learning [76]. In this chapter, we focus on transferring knowledge from a predictive model to obtain a bias-free estimation for the interventional distribution.

## 5.2 *do*-Calculus and Treatment Effect Calculation

To formalize the problem of effect restoration, we use probabilistic graphical models [44] and Pearl's *do*-calculus [74]. For example, in the graphical model shown in Figure 5.1a, we denote the direct effect of $X$ on $Y$ as $P(Y \mid do(X))$. Generally in observational studies, this quantity is different than simply conditioning on $X$ (i.e., $P(Y \mid X)$). Conditioning implies the probability distribution of $Y$ in each possible world defined by an observed $X$ value. However, the *do* operator implies actively setting the value of $X$ (i.e., intervention). Hence, $P(Y \mid do(X))$ represents the effect of actively manipulating the values of $X$ and $P(Y|X)$ represents passive observation. We refer the reader to the book by Pearl [74] for more details about *do*-calculus.

|  | Weight Gain |  |
|---|---|---|
| Activity Level | 0 | 1 |
| 0 | 0.20 | 0.80 |
| 1 | 0.60 | 0.40 |

Table 5.1: Effect of activity level on weight gain.

For the graphical model in Figure 5.1a, when all variables are fully observed and under identifiability conditions (again, see the book by Pearl [74] for details about such conditions), the probability $P(Y \mid do(X))$ can be estimated by:

$$P(Y \mid do(X = x)) = \sum_{U} \frac{P(X, Y, U)}{P(X \mid U)} \tag{5.1}$$

Given this *interventional* distribution, for a binary X, the *treatment effect* (TE) of $X$ on $Y$ can be calculated as [75, 72]:

$$TE = E(Y \mid do(X = 1)) - E(Y \mid do(X = 0)). \tag{5.2}$$

From perfect observations of $X$, $Y$, and $U$, consistent estimates of TE can be obtained using various modeling methods [37, 63, 73, 97]. However, these methods fail to provide consistent estimates if $U$ is measured with error [23, 92]. Kuroki et al. [47] recently extended the *do*-calculus to adjust for confounding variables with measurement error. Figure 5.1b shows the graphical model structure they assume in their extended framework. They propose that under certain conditions, the *TE* of $X$ on $Y$ can be restored bias-free when a proxy variable for the confounding $U$ (i.e., $W$) is observed and the error distribution (i.e., $P(W \mid U)$) is known. (See work by Kuroki et al. [47] for further details.)

**Treatment Effect Calculation**

Our experiments use a specific formulation for the $TE$ estimator defined by Kuroki et al. [47] with the *do*-calculus framework. Illustrating with our earlier example of activity level, weight gain, and age, our goal is to estimate the interventional distribution of $P(WeightGain|do(ActivityLevel))$. Let us assume the interventional distribution shown in Table 5.1.

According to this distribution, when a person has low activity level, the odds of weight gain is $\frac{0.80}{0.20} = 4$. Whereas, when a person has high activity level, the odds of weight gain is $\frac{0.40}{0.60} = 0.66$. We define the difference in the log-odds ratio for different treatment values as the causal effect of treatment (e.g., activity level) on outcome (e.g., weight gain). Mathematically, the causal effect of a binary treatment variable on a binary outcome variable (referred as *TE*) is:

$$TE = \log\left(\frac{P(Y_1 \mid do(X_1))}{P(Y_0 \mid do(X_1))}\right) - \log\left(\frac{P(Y_1 \mid do(X_0))}{P(Y_0 \mid do(X_0))}\right)$$

## 5.3 Effect Restoration and Underlying Structure

The effect restoration method proposed by Kuroki et al. [47] assumes that a confounding bias exists as shown in the graphical structure in the first column of Figure 5.2. Here, we relax this assumption and characterize the performance of effect restoration under all possible modifications of the simple graphical structure suggested in the original paper.

Figure 5.2 shows all possible modified structures between X, Y, W, and U with the following assumptions, which are typical in many causal inference studies in social and medical sciences.

**Assumption 1.** $X$ is temporally prior to $Y$.

**Assumption 2.** $W$ is a noisy measurement of $U$.

**Assumption 3.** $U$ is temporally prior to $X$ and $Y$.

We employ simulation studies to characterize the performance of effect restoration for each possible graphical model structure.

### 5.3.1 Data Generation

We generated discrete and continuous synthetic data consistent with the graphical structures in Figure 5.2. We explain the discrete data generation process for the graphical structure shown in the

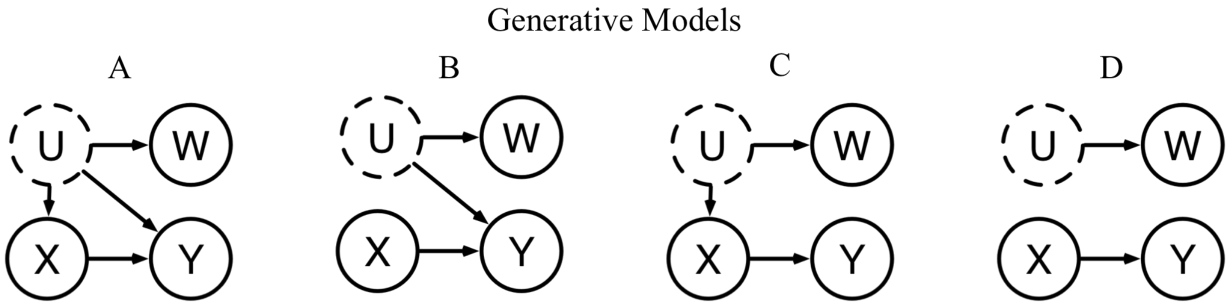


Figure 5.2: All possible underlying structures with temporal ordering assumptions

first column of the table. Other graphical structures follow almost the same steps, except that the added, removed, or reversed dependencies in the structure entail adding, removing, or changing the order of steps in the data generation process, respectively.

In the generative processes below, italic letters denote scalar values (e.g., *N*); upper-case bold characters denote vectors (e.g., **W**); each element of a vector is accessed by an index subscript (e.g., $w_i$); correlations between variables are referred with subscripts such as $\rho_{uw}$ denoting the correlation between **U** and **W**; marginal and conditional probabilities are denoted by the upper-case letter *P*. Following are the steps to generate binary data for the structure in column A in Figure 5.2.

We vary $\rho_{uw}, \rho_{ux}, \rho_{uy}$ and $\rho_{xy}$ correlation values between (0,1). We model $P(Y \mid U, X)$ as a *noisy-or* conditional probability distribution by using the following formula as suggested in Kohler et al. [44]:

$$P(Y = 0 \,|\, \mathbf{U}, \mathbf{X}) = (1 - \lambda_0) * (1 - \rho_{uy})^{\mathbf{U}} * (1 - \rho_{xy})^{\mathbf{X}} \tag{5.3}$$

$$P(Y = 1 \,|\, \mathbf{U}, \mathbf{X}) = 1 - (1 - \lambda_0) * (1 - \rho_{uy})^{\mathbf{U}} * (1 - \rho_{xy})^{\mathbf{X}} \tag{5.4}$$

We set the value of $\lambda_0 = 0.01$ as the noise parameter of the noisy-or model. Note that we add a constant bias to values of W and the amount of bias is defined by the $\rho_{uw}$.

For each parameter setting, $\{\rho_{uw}, \rho_{uy}, \rho_{ux}, \rho_{xy}\}$, we generate 50 separate data sets. Each dataset includes 10,000 instances, where each instance is a list of values for $U$, $X$, $Y$, $W$. As noted, for other graphical structures we revise the data generation process for the corresponding removed

```
1. Set correlation values, $\rho_{uw}$, $\rho_{uy}$, $\rho_{ux}$, and $\rho_{xy}$.
2. Draw a prior for $P_u \sim Uniform(0,1)$.
3. Draw $N$ values for $\mathbf{U}$ from $Bernoulli(P_u)$.
4. For each $u_i$ in $\mathbf{U}$:
    (a) Draw a $p' \sim Uniform(0,1)$.
    (b) If $p' < \rho_{uw}$, set $w_i = u_i$;
        else draw a value for $w_i \sim Bernoulli(p = 0.8)$.
5. For each $u_i$ in $\mathbf{U}$:
    (a) Draw a $p'$ from $Uniform(0,1)$.
    (b) If $p' < \rho_{ux}$, set $x_i = u_i$;
        else draw a value for $x_i \sim Bernoulli(p = 0.5)$.
6. Draw $\mathbf{Y} \sim Binomial(N, P(Y = 1 \mid \mathbf{U}, \mathbf{X}))$
```

dependency. For example, in the structure in Figure 5.2 column B, the dependency between $U$ and $X$ is removed. To account for this in our data generation process, we sample a prior for $P_X$ from a uniform distribution; we sample values for $X$ using the prior instead of using $P(X \mid U)$.

In our experiments, we calculate the true treatment effect (i.e., TE) as our ground truth by using the values of $U$. We estimate the treatment effect (i.e., TE') with the following three approaches: (1) *Ignore W*: Simply ignoring the measurements of $W$, (2) *Ignore measurement error:* Using $W$ and ignoring the measurement error, and (3) *Effect restoration:* Using $W$ and adjusting for the measurement error. Note that these three methods do not use values of $U$ (i.e., they can only use values of $X$, $Y$ and $W$). We measure the bias for each approach in each experiment by normalized error:

$$\epsilon = \frac{TE - TE'}{TE}$$

We measure the standard deviation of error values across experiments.

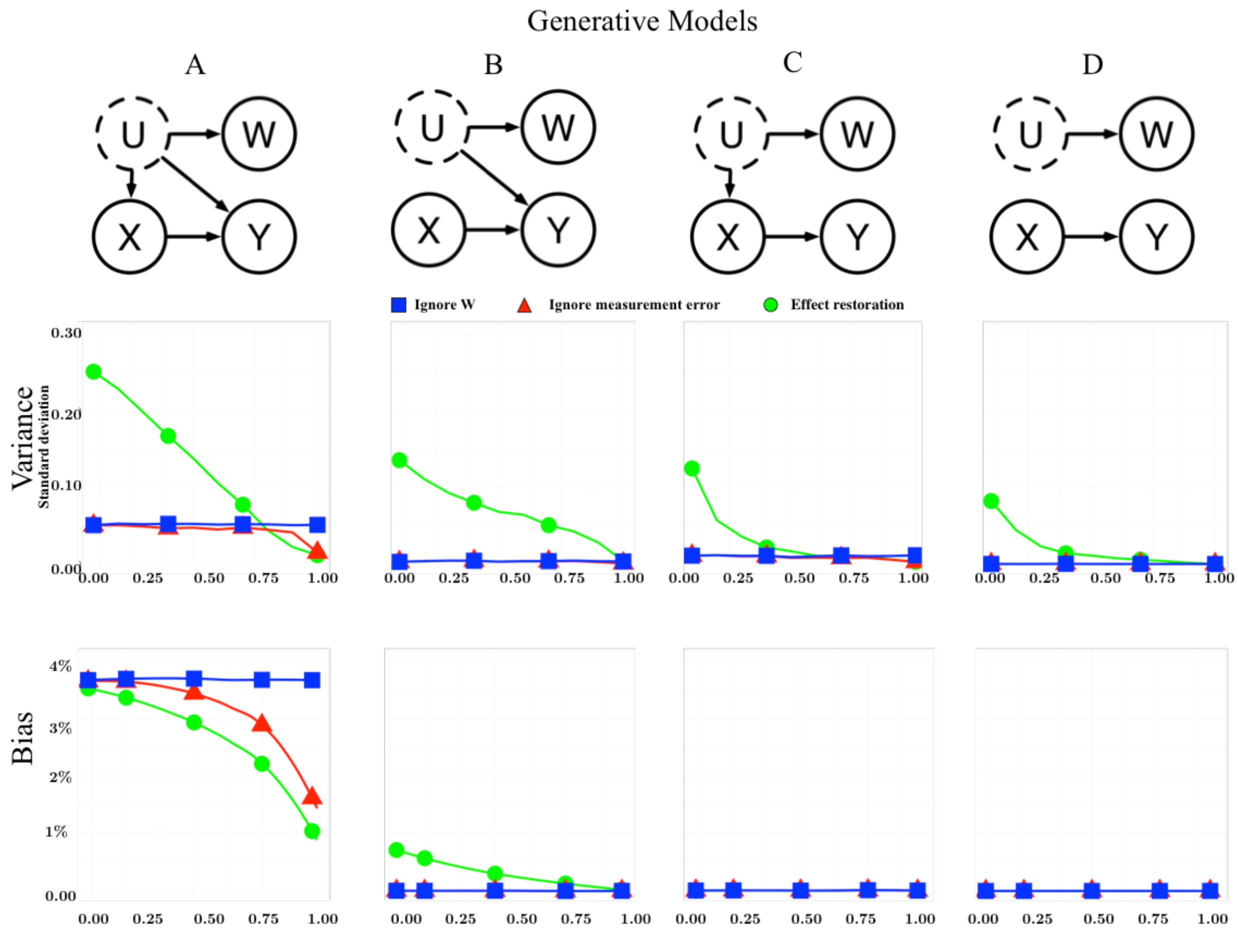


Figure 5.3: Variance and bias results for different underlying structures

### 5.3.2 Bias and Variance for Different Underlying Structures

We perform simulation analysis for each of the graphical structure in the first row of Figure 5.3. In the second and third row, we show both the variance and the bias in estimating the treatment effect, respectively. In each plot, we show the behavior for each of the three different approaches as the measurement error changes along the $x$-axis. We calculate the measurement error by the strength of dependency between $U$ and $W$ using the Cramer's V $\phi$ coefficient[2]. The stronger the dependence, the weaker the measurement error. On the $y$-axis we plot locally smoothed normalized

[2] www.en.wikipedia.org/wiki/phi_coefficient

error for the bias and locally smoothed standard deviation for the variance for the three different approaches. Figure 5.3 shows the results for a fixed treatment and confounding effect for all graphical structures considered.

We observe that, for the graphical model structure in the first column in Figure 5.3, when the confounding variable is simply ignored, unsurprisingly, we see a constant bias in our estimate of the treatment effect. However, when values of $W$ are used as if they are the perfect observations of $U$ (i.e., when the measurement error is ignored), the bias in treatment effect estimate is reduced. Unsurprisingly, this reduction is especially significant, when values of $W$ is highly correlated with values of $U$ (i.e., measurement error is small). Finally, when values of $W$ is used with the effect restoration adjustment, the bias is consistently reduced more than the other approaches.

Furthermore, the smaller the measurement error between $W$ and $U$, the smaller the bias in estimation. Also, the larger the measurement error, the more the relative benefit of explicitly adjusting for it using effect restoration versus simply ignoring it. However, when $U$ and $W$ are poorly correlated (i.e., measurement error is high), applying effect restoration comes with the cost of increased variance, as shown in the second row in Figure 5.3. In a way, when effect restoration is used with poorly correlated proxy variables, it transforms the problem of estimating treatment effect from a high-biased one to a high-variance one.

For the other graphical model structures in Figure 5.3 columns B, C, D, we observe that ignoring or using $W$ directly provide consistent estimates for the treatment effect, however, explicitly applying effect restoration might increase bias as well as variance. Hence, applying effect restoration regardless of the underlying structure can result in an incorrect estimate.

We perform additional experiments for the graphical model structure in column A by changing the values of treatment effect and confounding effect. Figure 5.4 shows the results of our experiments. In these plots, the treatment effect increases along the big y-axis and confounding effect increases along the big $x$-axis. Along the $x$-axis in each plot, the strength of effect between $U$ and $W$ increases. These results suggest that the effect restoration is most effective when the treatment effect is small and the confounding effect is high.

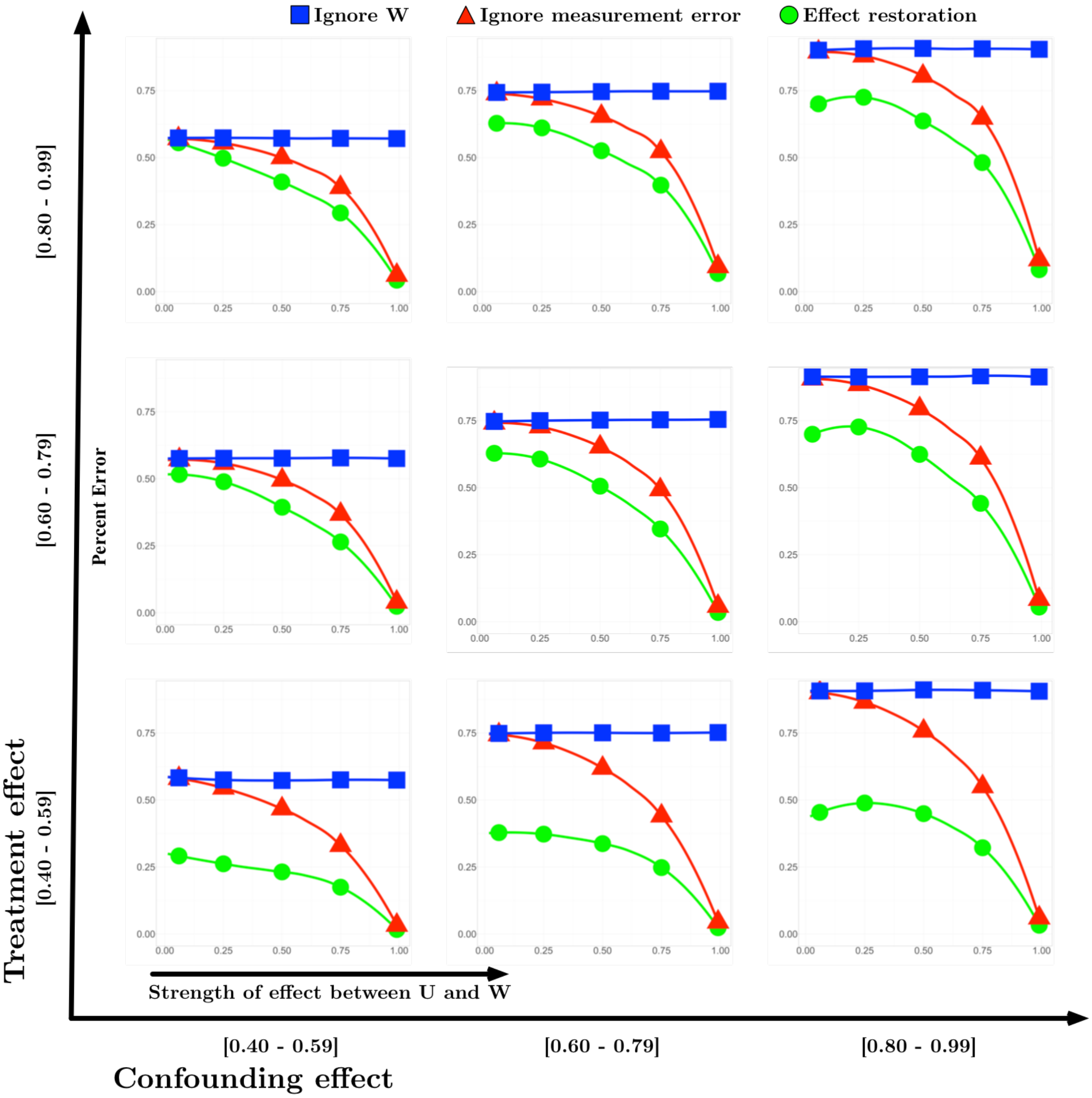


Figure 5.4: Bias and variance as the treatment and confounding effects change

We have performed the same experiments with continuous data, and we reach the same conclusions in those experiments. See Appendix Section A.2 for the results of the continuous experiments.

## 5.4 Detecting the Underlying Graphical Structure

In the previous section, we presented empirical evidence that when the underlying structure deviates from the confounding variable case, the adjustment provided by effect restoration can increase bias and variance. This raises a natural question: *Can we detect when to apply effect restoration?* Instead of assuming that the confounding bias exists, we propose to verify if it exists by using $d$-separation and typical temporal ordering constraints on the variables.

Note that $U$ may or may not be a confounding variable for $X$ and $Y$. Our goal is to identify sufficient conditions to determine if $U$ is a confounding variable and only apply effect restoration when it is.

### Possible Underlying Structures

In Table 5.2, we list all possible underlying structures with variables $X$, $Y$, $W$, and $U$ that satisfy the stated assumptions 1 and 2 in Section 5.3. There are nine possible graphical structures. We individually account for dependence and independence relationships between variables in each of the possible structures. One of the possible structures contains a cycle (i.e., the structure in the last column of Table 5.2 is not a directed acyclic graph.) and hence out of the scope of our discussion.

In four of these structures, $U$ is temporally prior to $X$ and $Y$ (i.e., columns A through D). In the remaining four structures, $U$ is temporally posterior to $X$ and $Y$ (i.e., columns E to H). In the rows of Table 5.2, for each graphical model structure, we list all the marginal and conditional independence relations between $X$, $Y$, and $W$ using *d-separation* criterion. In the columns of the table, we list the possible underlying graphical models. Each column vector in the table corresponds to the expected conditional dependence and independence relations for the corresponding graphical model.

We make several observations. First, we note that some of these graphical structures are perfectly distinguishable from conditional independence relations (i.e., structures in columns B, D, and F in Table 5.2). These structures can be identified by analyzing observed values of $X$, $Y$,

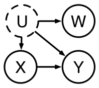
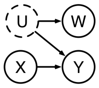
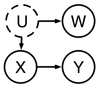
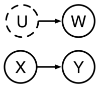
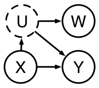
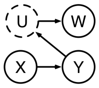
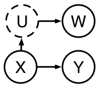
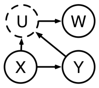
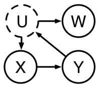
| | U is temporally prior to X and Y | | | | U is temporally posterior to X and Y | | | | Cycle |
|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | I |
| |  |  |  |  |  |  |  |  |  |
| $X \perp\!\!\!\perp Y$ | ✖ | ✖ | ✖ | ✖ | ✖ | ✖ | ✖ | ✖ | N/A |
| $X \perp\!\!\!\perp Y\|W$ | ✖ | ✖ | ✖ | ✖ | ✖ | ✖ | ✖ | ✖ | N/A |
| $X \perp\!\!\!\perp W$ | ✖ | ✔ | ✖ | ✔ | ✖ | ✖ | ✖ | ✖ | N/A |
| $X \perp\!\!\!\perp W\|Y$ | ✖ | ✖ | ✖ | ✔ | ✖ | ✔ | ✖ | ✖ | N/A |
| $Y \perp\!\!\!\perp W$ | ✖ | ✖ | ✖ | ✔ | ✖ | ✖ | ✖ | ✖ | N/A |
| $Y \perp\!\!\!\perp W\|X$ | ✖ | ✖ | ✔ | ✔ | ✖ | ✖ | ✔ | ✖ | N/A |

Table 5.2: Conditional (in)dependence relationships for all simple graphical structures.

and $W$, assuming accurate conditional independence tests. However, some structures are indistinguishable with the given set of conditional independence relations (e.g., structures in A, E, and H share exactly the same set of conditional independence relations).

Second, if we also make a common assumption that covariates are measured *pre-treatment* [10, 93, 100, 107], then only the structures in columns of A, B, C, and D will be possible. Furthermore, conditional independence relations would be sufficient to distinguish among these graphical model structures.

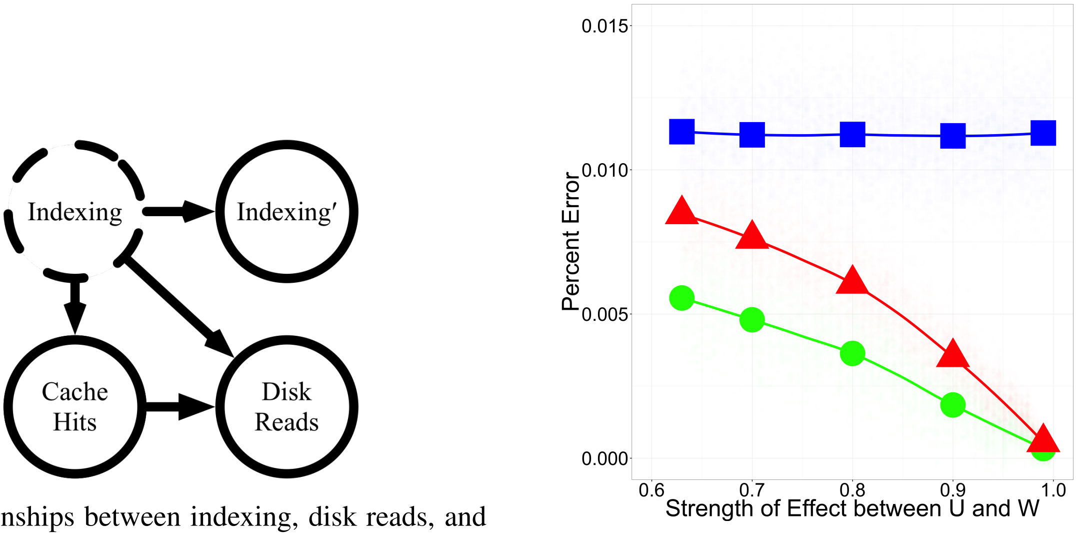
Thus, common temporal assumptions and conditional independence relations are sufficient to determine the underlying graphical structure from $X$, $Y$, and $W$.

## 5.5 Effect Restoration with Real Data

In this section, we evaluate the performance of effect restoration on causal distributions obtained from real-world settings. We use the experimental data compiled by Garant et al. [33] about the effects of interventions on large-scale software systems.

Specifically, we use their experimental data about *PostgreSQL*[3], a large open-source relational database management system. The authors ran over $11,000$ queries under each of eight different

[3] https://www.postgresql.org



(a) Relationships between indexing, disk reads, and cache hits on a Postgress database system.

(b) Bias in cache hits effect on disk reads estimation.

Figure 5.5: Effect restoration with real experimental data.

system configurations. This creates a nearly ideal data set for causal estimation because each subject (i.e., query) is observed in each treatment condition (i.e., configuration setting). This approach allows direct interventional estimates of the effect of treatment variables on outcome variables in the context of a large number of other covariate variables representing characteristics of queries and intermediate states of the database server.

The authors then use *GES*, a score-based algorithm for structure learning [24], to recover the underlying structure for the PostgreSQL domain. From their partially learned graphical structure, we focus on specific variables in which the sub-structures satisfy the effect restoration conditions as shown in Table 5.2.

We identify two cases in which we can apply effect restoration. One is shown in Figure 5.5a and represents the relationships between *indexing*, *disk reads*, and *cache hits*. Generally speaking, indexing affects both disk reads and cache hits. Indexing reduces the disk reads (i.e., a result of a query can be retrieved with fewer disk block reads) and increases the cache hits. Figure 5.5a shows the cache hits as a cause of disk reads. The authors note that the direction of this edge between disk reads and cache hits might also be in the opposite direction. In our analysis, we separately analyzed graphical structures in both directions and the results for effect restoration are similar in both cases.

For simplicity, we converted each variable to a binary variable by using the median value of each variable as a threshold. Our goal is to estimate the effect of cache hits on disk reads under noisy measurements of indexing. To obtain such noisy measurements, we manually added noise to the values of indexing by overwriting the values for a subset in indexing with results coming from a random sampling process.

Then we use these noisy measurements to adjust for the confounding bias of indexing. Similar to our previous experiments, we compare three approaches: (1) Ignoring the values of noisy measurements for indexing, (2) Using the values of indexing, ignoring that they are noisy, (3) Using the values by correcting for the noise in their measurements. We calculate the true effect by using the true values of indexing.

In Figure 5.5b, we plot the normalized error in the estimated treatment effect with respect to measurement noise in indexing. Similar to our results on simulated data, estimation with correction provides significantly smaller bias than alternatives. In addition, as the measurement error of the confounding variable increases (i.e., the strength of effect between $U$ and $W$ decreases), the relative benefit of applying measurement error correction increases over simply ignoring the measurement error.

## 5.6 Effect Restoration with Predictive Models

Adjusting for a confounding variable by using noisy measurements can also be thought as adjusting for an unobserved variable when both its predictions and the error distribution of such predictions can be inferred using an independent external estimation process. For example, assume we want to estimate the effect of *activity level* on *weight gain*. Assume *age* is a confounding variable (i.e., age is related to both activity level and weight gain) and that it is unobserved for the population of interest. Clearly, in this hypothetical example, we need to adjust for the confounding bias of age to get a consistent estimate of activity level on weight gain.

One idea might be to estimate age by using other observables for the subjects under study. For example, such observables might be based on users' social media content [66, 77], links on social networks [110], first names [68], or names combined with personal images [32].

We can use the predictions of such models as noisy measurements of age and along with their error distributions adjust for its confounding bias. This idea assumes that the population under study is similar to the population from which the predictive model was trained so that we can transfer the knowledge of that model to obtain the corresponding predictions and error distribution. This assumption, referred to as external validity, is common to nearly all statistical causal modeling.

Here, a priori, we hypothesize that *age* might be a confounding variable, and we employ a latent variable model to estimate its values. Similarly, if there are other confounder variables such as ethnicity or income, we can separately use latent variable models to adjust for their effects as long as the confounder variables are independent of each other.

We evaluate the idea of using predictive models for effect restoration by constructing a scenario in which we observe activity levels, and weight gain recordings of a population along with their first names. Although one can use more complicated models for predicting age, here we use the model introduced in Chapter 4 due to its simplicity.

In our experiments, we generate synthetic data with activity levels, weight gain, and the first name for each subject. We then use first names to infer an age value for each subject. Finally, we use the inferred age values along with the corresponding error distribution to adjust for the confounding bias due to age.

One might suggest that instead of using a model to estimate age, we could simply adjust for the values of first names. There are three reasons for avoiding this approach. First, the number of possible first names is very large and using them directly would lead to a high-variance estimator of causal effect for most data sets.

Second, conditioning on first names leaves the back-door path unblocked between activity levels and weight gain [47], as shown on in Figure 5.6. (See [74] for a detailed discussion of back-

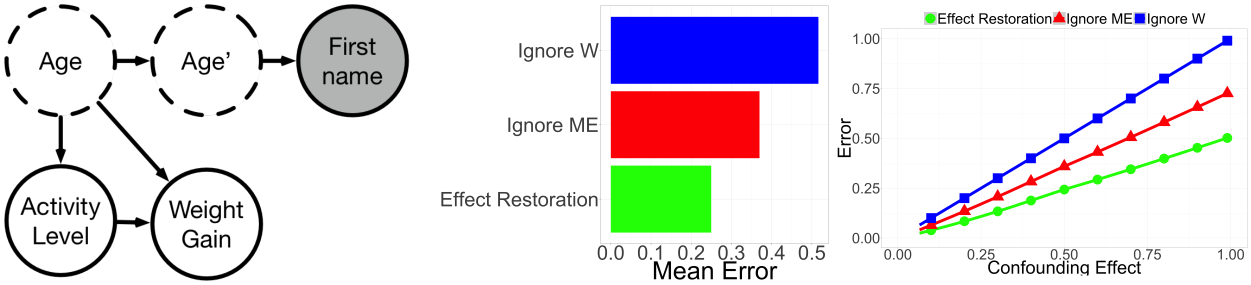


Figure 5.6: Effect restoration for unobserved variables with predictive models

door paths.) This implies that the confounding bias would still exist. Third, our proposal is to plug in *any* predictive model for unobserved confounders, and such models can use many independent variables rather than just one. Again, conditioning on many independent variables can lead to a high-variance estimator. In fact, from the perspective of effect restoration adjustment, using high-capacity models in causal estimation is desirable to drive down prediction error and subsequently reduce bias.

Our proposed approach—using a predictive model for effect restoration in pursuit of a low-bias estimator of causal effect—is similar in spirit to the use of predictive models for propensity score matching [88]. Both approaches use a predictive model to summarize the effect of a potentially large number of variables. Propensity score matching aims to summarize all observed covariates, however, we focus on adjusting for unobserved confounding variables.

Figure 5.6 shows the graphical model representation of the experiment. As before, we compare estimating the effect when a confounding variable is ignored, when the error in the confounding variable is ignored, and finally when effect restoration is used. Our results with an independent estimation process are similar to those obtained earlier from both the simulation and real data experiments. We observe the most reduction in bias when correction based on measurement error is used. We also show that, as the influence of confounding variable increases, the relative benefit of effect restoration increases.

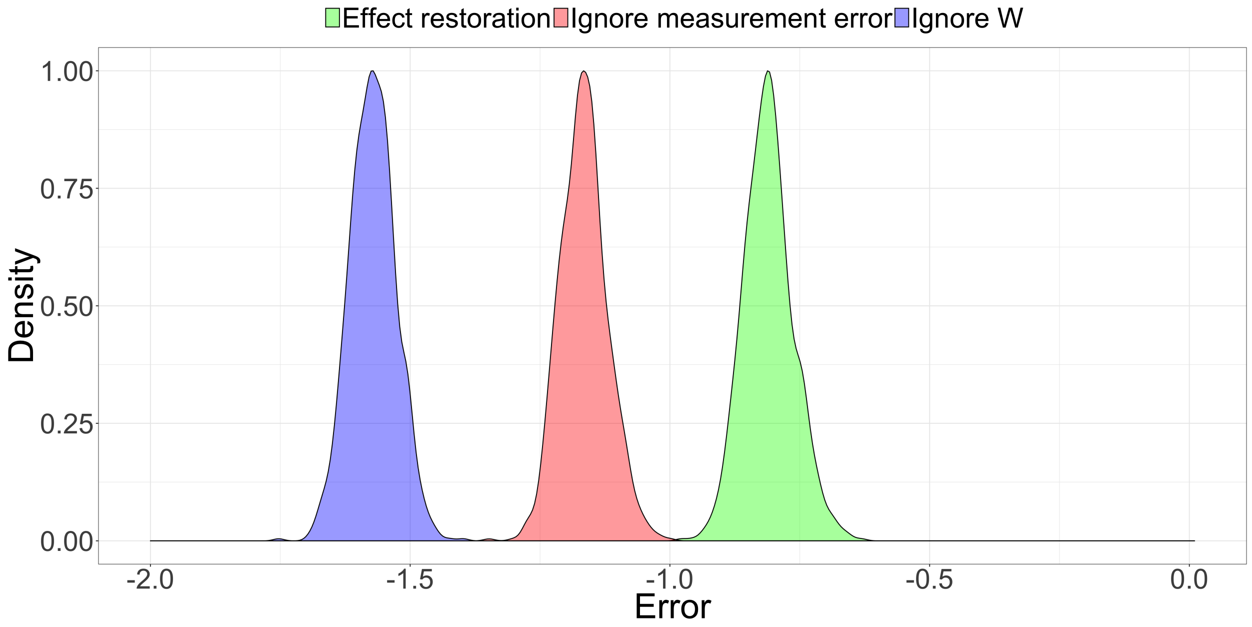


Figure 5.7: Density plot of estimations using three alternative models. Using effect restoration reduces the bias practically in *all* cases. Similarly, using the information in the proxy variables reduces the bias practically in *all* cases over simply ignoring it.

### Confidence intervals with bootstrapping

We perform confidence interval estimation by performing bootstrapping. Specifically, for a fixed confounding and treatment effect, we perform estimation of the treatment effect 1000 times. In each iteration, we sample 10K instances for activity level, weight gain, and first name. We calculate error for each of the three methods. We use the treatment effect estimations from all iterations to empirically estimate density distributions for error values corresponding to the three separate methods.

Figure 5.7 shows the density plots for the error values corresponding to the three methods: (1) Ignore W, (2) Ignore measurement error, (3) Effect restoration. Practically, in all iterations using effect restoration reduces the bias more than ignoring the measurement error. The empirical density distributions of the error values intersect at -0.98. The probability that an estimate from effect restoration mechanism will have an error value less than -0.98 is practically 0. Similarly, probability that an estimate from ignoring measurement error will have an error value greater than -0.98 is also practically 0.

Additionally, when we compare the density of the error values between simply ignoring the proxy variable and using the proxy variable (but ignoring the measurement error), we find that using the proxy variable reduces the bias practically in all iterations.

## 5.7 Conclusions

In this chapter, we characterize the behavior of effect restoration with measurement error under all possible simple graphical structures as suggested by Kuroki et al. [47]. Unsurprisingly, we show that it is desirable to use effect restoration only in one of the four possible graphical models. According to our simulation analysis, effect restoration adjustment is most effective for small treatment and large confounding effects. Furthermore, we show that with common temporal assumptions among variables and simple d-separation rules, we can identify if the underlying structure matches the desirable one. We also show empirical evidence that effect restoration adjustment can reduce bias on causal estimation tasks in real data. Finally, we show that this mechanism can be used to account for unobserved confounding bias when used with independent predictive models and their corresponding error distributions.

# CHAPTER 6

# CONCLUSIONS AND FUTURE WORK

Causal estimation from observational data has been an important problem for researchers in diverse fields and has become widespread with the availability of large and rich datasets. This thesis focuses on causal estimation when latent common cause variables exist. We present a mechanism to adjust for the effects of such latent variables using predictive models. Specifically, first, we employ predictive models to estimate the values of confounder variables. Second, we use such inferred values as the proxy variables of unobserved confounders with one caveat that the proxy variables are measured with error. We use an effect restoration model based on graphical models as our measurement error model to deal with this estimation error, and adjust for the confounding effect due to unobserved variables. We highlight this mechanism as a novel method to remove the bias in causal estimation due to important unobserved confounder variables.

We have presented new applications of causal estimation methods to social media domains, new predictive models for key demographic variables, and a new methodology to deal with latent common cause variables using predictive models. In this final chapter, we discuss concluding remarks and identify future research directions.

## 6.1 A Novel and Early Application of QEDs to Data About Social Media Platforms

We present one of the earliest applications of QEDs, a set of causal estimation methods highly used by social scientists, to an arguably popular social media platform, i.e., the Stack Overflow website. We illustrate three basic QEDs and use them to answer causal questions about the platform.

Our results show no significant effect of having an already posted high-quality answer on the number of subsequent answers posted by other users. Furthermore, specific badges designed to drive engagement seem to work until the badge is received, however, engagement is significantly reduced after. This suggests that an alternative badge mechanism where tiered-badges with increasing level of exclusivity to help sustain contributions from users. Finally, we show that answer order has no effect on the number of up-votes, contrary to the popular perception among the users. This suggests that users might be taking their time to read all answers regardless of their relative order.

We also discuss the assumptions and limitations of QEDs, specifically, their behavior when latent common cause variables exist. We propose overcoming confounding bias due to latent common cause variables by using estimates from predictive models.

## 6.2 A New Predictive Model with Distant Supervision for Ordinal Variables

We develop a novel predictive model based on generative models to predict ordinal latent variables with distant supervision. We use this model to predict age using first names in a given population. Our model is distantly-supervised in the sense that it uses Social Security baby name data for each year as a prior and is capable of both providing individual-level predictions as well as population-level predictions. We show the benefit of explicitly modeling the ordinal dependency among age groups.

We apply our model to estimate demographic information about US Twitter users. First, we closely replicate a Pew Research report about the active age groups on Twitter, finding that *18-29* age group is the most active. Legally limited by the user groups they could reach by phone surveys, Pew Research report age groups older than *18* (i.e., they cannot legally survey users younger than *18*). Our method is free from such a limitation and we find that less than *18*-year-old is the second most active age group on Twitter.

Our results also indicate that users older than *50* are under-represented; conversely, users in the *18-29* age group are over-represented compared to their respective presence in the internet users population. We identify times in a day each age group is active on Twitter, finding that users in *65+* age group and *18-29* age group are active on Twitter at completely opposite times during the day. We show evidence that different age groups engage in different topics and follow different users. Finally, we find that young and senior Twitter users exhibit high-order assortative mixing in their follow relationships, whereas middle age users show low assortative mixing.

We develop a simple yet effective predictive model for age, a traditional demographic variable, by using a generative model. We propose that for causal questions where age might be a latent variable, estimations from predictive models such as presented in this thesis can help adjust for its effect. Since every estimation process inherently includes error, we point that such estimations would be noisy and we characterize potential ways to deal with such noise.

## 6.3 Characterization of Effect Restoration with Measurement Error and Its Application with Predictive Models

We characterize the behavior of effect restoration with measurement error under all possible scenarios a treatment variable, an outcome variable, a latent variable, and its proxy variable can form. Employing simulation studies, we show that only when the latent variable is a common cause the effect restoration can reduce bias. In all other scenarios, applying effect restoration can increase variance and slightly increases bias.

In the common cause case, we show that effect restoration is most effective for small treatment and large confounding effects. Using theoretical analysis, we show that simple d-separation rules along with basic temporal assumptions are sufficient to detect which underlying graphical model holds given a treatment, an outcome, and a proxy variable, enabling to decide whether effect restoration should be applied.

Furthermore, we show that effect restoration reduces bias on real data sets obtained from randomized experiments of software systems. Finally, we combine effect restoration with estimations of a latent common cause variable obtained from a predictive model to reduce bias in causal estimation. We argue that data from social media platforms can be enriched with latent variable models to account for the confounding effects of traditional demographic variables such as age, ethnicity, and income. For each unobserved confounder variable, first, we employ a predictive model to estimate its values. Second, we use such inferred values as the proxy variables of unobserved confounders with one caveat that the proxy variables are measured with error. We use an effect restoration model based on graphical models to account for the measurement error. We can adjust for multiple unobserved confounder variables using separate predictive models as long as each confounder variable is independent of each other.

There are several future research directions. First, the proposed effect restoration method can be generalized for mixed-type data sets. Second, the use of high-capacity predictive models and their limitations can be explored. Third, existing causal discovery algorithms can be extended to account for measurement errors using an effect restoration mechanism. Fourth, instead of independently inferring values for a latent common cause variable and estimating a treatment effect, both of these processes can jointly be modeled and inferred.

# APPENDIX

# SUPPLEMENTAL MATERIALS

## A.1 Keywords Used to Filter Tweets for Corresponding Topics

Here we list the keywords we use to filter tweets related to the corresponding topic.

**Immigration Act**–*#cir OR #immigration OR #CIR OR #immyouth OR #DREAMact OR #cirasap OR #dwn OR #StopICE*

**Fiscal Cliff**–*(”fiscal cliff” OR fiscalcliff)*

**Gun Control**–*(obama OR romney) AND ”gun control”*

**Blizzard Nemo**–*(”winter storm” OR blizzard OR Nemo OR winterstorm OR (snow AND storm) OR snowstorm OR snow*

**Boston Strong**–*BostonStrong OR ”Boston Strong” OR OneFundBoston OR ”One Fund” OR ”Boston Marathon” OR BostonMarathon OR WeAreBoston OR BostonStrongest OR BelieveInBoston OR WeAreOneBoston OR PrayForBoston*

**NBA Finals**–*(”San Antonio” OR spurs OR ”Tim Duncan” OR ”Tony Parker” OR Miami OR ”the heat” OR ”Dwayne Wade” OR LeBron OR MIA OR SA) AND (finals OR NBA finals OR win OR winning OR champion OR champions OR ring OR ”win finals”) AND -(eastern OR east OR west OR western OR Indiana OR Pacers OR ”game 7” OR Memphis OR Grizzlies OR HTTP OR ”moving on”)*

## A.2 Effect Restoration with Continuous Data

We applied effect restoration framework to continuous data generated from the simple graphical models, as shown in the first row of Figure A.1.

### A.2.1 Data Generation

Following steps describe the data generation process for the variables **X**, **Y**, **W**, **U** when **U** is a confounder for **X** and **Y**. $\epsilon_{..}$ corresponds to the Gaussian error.

1. $U \sim Normal(\mu_u, \sigma_u)$
2. $W \sim \beta_{uw} U + \epsilon_{uw}$
3. $X \sim \beta_{ux} U + \epsilon_{ux}$
4. $Y \sim \beta_{uy} U + \beta_{xy} X + \epsilon_{.y}$
5. $\epsilon_{..} \sim Normal(0, \sigma_{..})$

In our experiments, we set $\mu_u = 0$, $\sigma_u = 1$. For the graphical model structure that corresponds to common cause scenario, we set $\beta_{uw} = 1, \beta_{ux} = 1, \beta_{uy} = 1$, and $\beta_{xy} = 2$. We systematically change the values of $\sigma_{..}$ between $0.05$ to $2$.

Figure A.1 summarizes the bias and variance when effect restoration is used with continuous synthetic data for the four graphical models. Similar to the results from simulations with discrete data, unsurprisingly, effect restoration reduces bias only when U is a confounder variable, and it comes with the cost of increased variance. Furthermore, it is worse to use effect restoration in other simple scenarios as it might increase bias and variance.

We further perform simulations for the confounder variable scenario by varying the error distributions to change the confounding effect. As shown in Figure A.2, these experiments show that the relative benefit of applying effect restoration is high when the confounding effect is high.

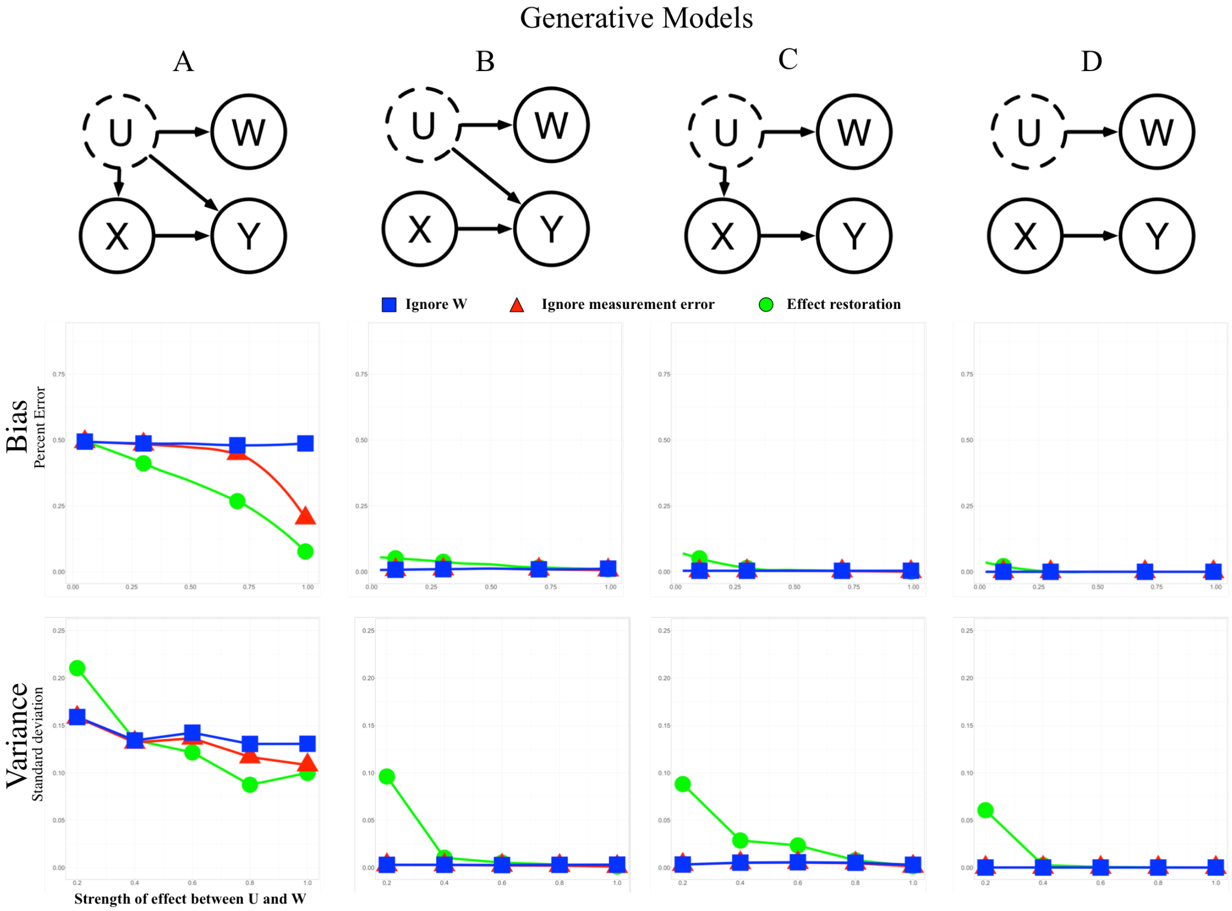


Figure A.1: Bias and variance for the graphical models identified in continuous experiments.

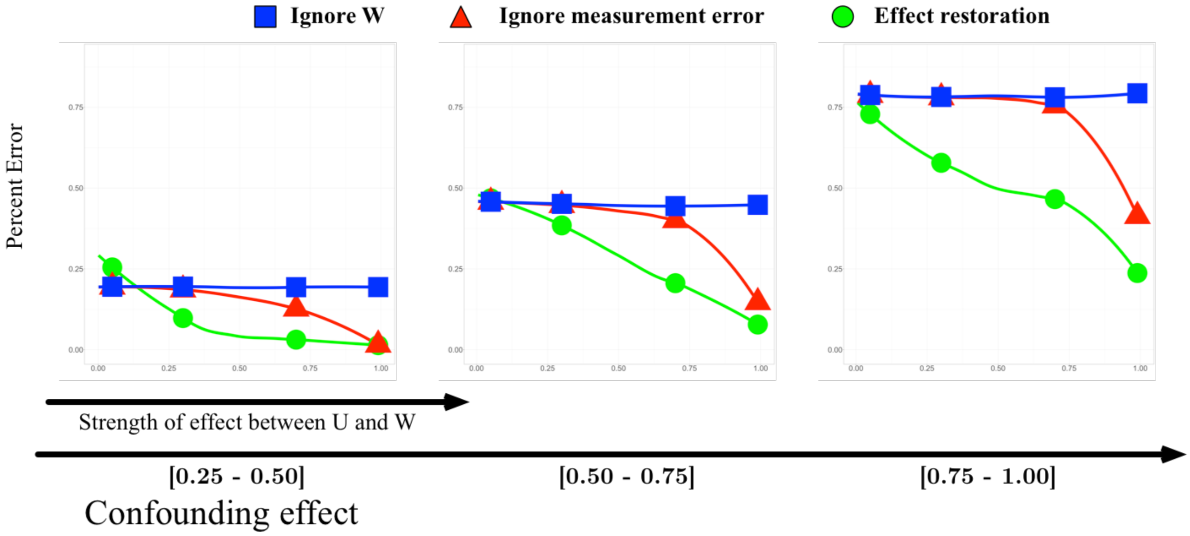


Figure A.2: Performance as the confounding effect increases.

# BIBLIOGRAPHY

[1] Adamic, Lada A, Zhang, Jun, Bakshy, Eytan, and Ackerman, Mark S. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web* (2008), ACM, pp. 665–674.

[2] Agrawal, Divyakant, Das, Sudipto, and El Abbadi, Amr. Big data and cloud computing: current state and future opportunities. In *Proceedings of the 14th International Conference on Extending Database Technology* (2011), ACM, pp. 530–533.

[3] Agresti, A. *Analysis of Ordinal Categorical Data*, second ed. Chapter 11 in Probability and Statistics. Wiley, 2010.

[4] Angrist, Joshua D., and Pischke, Jörn-Steffen. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Dec. 2008.

[5] Aral, Sinan, Muchnik, Lev, and Sundararajan, Arun. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences 106*, 51 (2009), 21544–21549.

[6] Aral, Sinan, Muchnik, Lev, and Sundararajan, Arun. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences 106*, 51 (Dec. 2009), 21544–21549.

[7] Aral, Sinan, and Nicolaides, Christos. Exercise contagion in a global social network. *Nature communications 8* (2017), 14753.

[8] Athey, S., and Imbens, G.W. The econometrics of randomized experimentsa. *Handbook of Economic Field Experiments 1* (2017), 73 – 140. Handbook of Field Experiments.

[9] Athey, Susan. Beyond prediction: Using big data for policy problems. *Science 355*, 6324 (Feb. 2017), 483–485.

[10] Athey, Susan, and Imbens, Guido W. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives 31*, 2 (May 2017), 3–32.

[11] Bakshy, Eytan, Karrer, Brian, and Adamic, Lada A. Social influence and the diffusion of user-created content. In *Proceedings of the 10th ACM Conference on Electronic Commerce* (New York, NY, USA, 2009), EC '09, ACM, pp. 325–334.

[12] Bengio, Yoshua. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning* (Bellevue, Washington, USA, 02 Jul 2012), Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham Taylor, and Daniel Silver, Eds., vol. 27 of *Proceedings of Machine Learning Research*, PMLR, pp. 17–36.

[13] Bentler, Peter M. Multivariate analysis with latent variables: Causal modeling. *Annual review of psychology 31*, 1 (1980), 419–456.

[14] Berk, Richard A. An introduction to sample selection bias in sociological data. *American Sociological Review 48*, 3 (1983), 386–398.

[15] Blei, David M., and Lafferty, John D. Correlated topic models. In *NIPS* (2005).

[16] Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *J. Mach. Learn. Res. 3* (Mar. 2003), 993–1022.

[17] Bollen, Johan, Gonçalves, Bruno, Ruan, Guangchen, and Mao, Huina. Happiness is assortative in online social networks. *Artificial life 17*, 3 (2011), 237–251.

[18] Bornfeld, Benny, and Rafaeli, Sheizaf. Gamifying with badges: A big data natural experiment on stack exchange. *First Monday 22*, 6 (2017).

[19] boyd, danah m., and Ellison, Nicole B. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication 13*, 1 (2007), 210–230.

[20] Campbell, D., and Stanley, J. *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, Chicago, IL, 1966.

[21] Carroll, Raymond J. *Measurement error in nonlinear models : a modern perspective*. Chapman & Hall/CRC, Boca Raton, 2006. Rev. ed. of: Measurement error in nonlinear models / R.J. Carroll, D. Ruppert, and L.A. Stefanski. 1998.

[22] Chang, Jonathan, Rosenn, Itamar, Backstrom, Lars, and Marlow, Cameron. epluribus: Ethnicity on social networks. In *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM-10)* (Washington DC, May 2010), AAAI Press, AAAI Press.

[23] Chesher, Andrew. The effect of measurement error. *Biometrika 78*, 3 (1991), 451–462.

[24] Chickering, David Maxwell, and Meek, Christopher. Finding optimal bayesian networks. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence* (San Francisco, CA, USA, 2002), UAI'02, Morgan Kaufmann Publishers Inc., pp. 94–102.

[25] Cohen, Patricia, Cohen, Jacob, Teresi, Jeanne, Marchi, Margaret, and Velez, C Noemi. Problems in the measurement of latent variables in structural equations causal models. *Applied Psychological Measurement 14*, 2 (1990), 183–196.

[26] Do, Chuong B, and Ng, Andrew Y. Transfer learning for text classification. In *Advances in Neural Information Processing Systems* (2006), pp. 299–306.

[27] Duggan, M., and Brenner, J. The demographics of social media users. *Pew Research Internet Reports* (February 14 2013).

[28] Fisher, Ronald A. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain 33* (1926), 505–513.

[29] Fisher, Ronald A. *Statistical Methods for Research Workers*. Oliver Boyd, Edinburgh, 1950.

[30] Friedman, Nir, Geiger, Dan, and Goldszmidt, Moises. Bayesian network classifiers. *Machine Learning 29*, 2 (Nov 1997), 131–163.

[31] Fuller, Wayne A. *Measurement error models*, vol. 305. John Wiley & Sons, 2009.

[32] Gallagher, A., and Chen, T. Estimating age, gender and identity using first name priors. In *Proc. CVPR* (2008).

[33] Garant, Dan, and Jensen, David. Evaluating causal models by comparing interventional distributions. In *The 2016 ACM SIGKDD Workshop on Causal Discovery* (2016).

[34] Goel, Sharad, Hofman, Jake M, and Sirer, M Irmak. Who does what on the web: A large-scale study of browsing behavior. In *ICWSM* (2012).

[35] Hirano, Keisuke, Imbens, Guido, and Ridder, Geert. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*, 4 (2003), 1161–1189.

[36] Iacus, Stefano M., King, Gary, and Porro, Giuseppe. Causal inference without balance checking: Coarsened exact matching. *Political Analysis* (2011).

[37] Imbens, Guido W., and Rubin, Donald B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, USA, 2015.

[38] Jensen, David, Neville, Jennifer, and Gallagher, Brian. Why collective inference improves relational classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2004), KDD '04, ACM, pp. 593–598.

[39] Jensen, Finn V. *Introduction to Bayesian Networks*, 1st ed. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.

[40] King, Gary. Ensuring the data rich future of the social sciences. *Science 331*, 11 February (2011 2011), 719–721.

[41] Kleinberg, Jon, Ludwig, Jens, Mullainathan, Sendhil, and Obermeyer, Ziad. Prediction policy problems. *The American economic review 105*, 5 (2015), 491–495.

[42] Kohavi, Ron, Henne, Randal M., and Sommerfield, Dan. Practical guide to controlled experiments on the web: Listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2007), KDD '07, ACM, pp. 959–967.

[43] Kollar, Daphne, and Friedman, Nir. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[44] Koller, Daphne, and Friedman, Nir. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[45] Krishnan, S Shunmuga, and Sitaraman, Ramesh K. Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs. In *Proceedings of the 2012 ACM Conference on Internet Measurement Conference* (2012), ACM, pp. 211–224.

[46] Kumar, Ravi, Lifshits, Yury, and Tomkins, Andrew. Evolution of two-sided markets. In *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010* (2010), pp. 311–320.

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

[48] Kusmierczyk, Tomasz, and Gomez-Rodriguez, Manuel. Harnessing natural experiments to quantify the causal effect of badges. *arXiv preprint arXiv:1707.08160* (2017).

[49] Lazer, David, Kennedy, Ryan, King, Gary, and Vespignani, Alessandro. The parable of google flu: traps in big data analysis. *Science 343*, 6176 (2014), 1203–1205.

[50] Lazer, David, Pentland, Alex, Adamic, Lada, Aral, Sinan, Barabási, Albert-László, Brewer, Devon, Christakis, Nicholas, Contractor, Noshir, Fowler, James, Gutmann, Myron, Jebara, Tony, King, Gary, Macy, Michael, Roy, Deb, and Van Alstyne, Marshall. Computational social science. *Science 323*, 5915 (2009), 721–723.

[51] Lerman, Kristina, and Galstyan, Aram. Analysis of social voting patterns on digg. In *Proceedings of the First Workshop on Online Social Networks* (New York, NY, USA, 2008), WOSN ’08, ACM, pp. 7–12.

[52] Lerman, Kristina, and Hogg, Tad. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th International Conference on World Wide Web* (New York, NY, USA, 2010), WWW ’10, ACM, pp. 621–630.

[53] Long, Mingsheng, Wang, Jianmin, Ding, Guiguang, Shen, Dou, and Yang, Qiang. Transfer learning with graph co-regularization. In *AAAI Conference on Artificial Intelligence* (2012).

[54] M., Aronow Peter, and A., Middleton Joel. A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference 1*, 1 (2013), 135–154.

[55] Maier, Marc, Marazopoulou, Katerina, Arbour, David, and Jensen, David. A sound and complete algorithm for learning causal models from relational data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence* (Bellevue, WA, July 2013), AUAI Press.

[56] Maier, Marc, Taylor, Brian, Oktay, Hüseyin, and Jensen, David. Learning causal models of relational domains. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (2010).

[57] Marlin, Benjamin M. Modeling user rating profiles for collaborative filtering. In *Advances in Neural Information Processing Systems 16*, S. Thrun, L. K. Saul, and P. B. Schölkopf, Eds. MIT Press, 2004, pp. 627–634.

[58] Martens, Edwin P, Pestman, Wiebe R, de Boer, Anthonius, Belitser, Svetlana V, and Klungel, Olaf H. Instrumental variables: application and limitations. *Epidemiology 17*, 3 (2006), 260–267.

[59] Maxwell, Scott E., and Delaney, Harold D. *Designing Experiments and Analyzing Data: A Model Comparison Perspective, Second Edition*, 2 ed. Routledge Academic, May 2003.

[60] McPherson, Miller, Smith-Lovin, Lynn, and Cook, James M. Birds of a feather: Homophily in social networks. *Annual review of sociology 27*, 1 (2001), 415–444.

[61] Mislove, Alan, Lehmann, Sune, Ahn, Yong-Yeol, Onnela, Jukka-Pekka, and Rosenquist, J. Niels. Understanding the Demographics of Twitter Users. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)* (Barcelona, Spain, July 2011).

[62] Mooij, Joris M., Stegle, Oliver, Janzing, Dominik, Zhang, Kun, and Schölkopf, Bernhard. Probabilistic latent variable models for distinguishing between cause and effect. In *NIPS* (2010), pp. 1687–1695.

[63] Morgan, S.L., and Winship, C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge University Press, 2007.

[64] Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[65] Newman, Mark EJ. Assortative mixing in networks. *Physical review letters 89*, 20 (2002), 208701.

[66] Nguyen, Dong, Gravel, Rilana, Trieschnigg, Dolf, and Meder, Theo. "How old do you think I am?": A study of language and age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (2013), ICWSM 2013.

[67] O'Connor, Brendan, Eisenstein, Jacob, Xing, Eric P., and Smith, Noah A. A mixture model of demographic lexical variation. In *Proceedings of NIPS Workshop on Machine Learning for Social Computing* (Vancouver, 2010).

[68] Oktay, Hüseyin, Ertem, Zeynep, and Firat, Aykut. Demographic breakdown of Twitter users: An analysis based on names. In *The sixth ASE International Conference on Social Computing* (May 2014), ASE.

[69] Oktay, Hüseyin, Taylor, Brian J., and Jensen, David. Causal discovery in social media using quasi-experimental designs. In *Proceedings of the SIGKDD/ACM Workshop on Social Media Analytics* (2010).

[70] Pan, Sinno Jialin, and Yang, Qiang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng. 22*, 10 (Oct. 2010), 1345–1359.

[71] Pearl, J. On Measurement Bias in Causal Inference. *ArXiv e-prints* (Mar. 2012).

[72] Pearl, Judea. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (San Francisco, CA, USA, 2001), UAI'01, Morgan Kaufmann Publishers Inc., pp. 411–420.

[73] Pearl, Judea. *Causality: Models, Reasoning and Inference*, 2nd ed. Cambridge University Press, New York, NY, USA, 2009.

[74] Pearl, Judea. *Causality: Models, Reasoning and Inference*, 2nd ed. Cambridge University Press, New York, NY, USA, 2009.

[75] Pearl, Judea. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence* (Arlington, Virginia, United States, 2012), UAI'12, AUAI Press, pp. 3–11.

[76] Pearl, Judea, and Bareinboim, Elias. Transportability of causal and statistical relations: a formal approach. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence* (2011), AAAI Press, pp. 247–254.

[77] Peersman, Claudia, Daelemans, Walter, and Van Vaerenbergh, Leona. Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents* (New York, NY, USA, 2011), SMUC '11, ACM, pp. 37–44.

[78] Pennacchiotti, Marco, and Popescu, Ana-Maria. A Machine Learning Approach to Twitter User Classification. In *International AAAI Conference on Weblogs and Social Media* (2011).

[79] Peysakhovich, Alexander, and Lada, Akos. Combining observational and experimental data to find heterogeneous treatment effects. *CoRR abs/1611.02385* (2016).

[80] Pourhoseingholi, Mohamad Amin, Baghestani, Ahmad Reza, and Vahedi, Mohsen. How to control confounding effects by statistical analysis. *Gastroenterology and Hepatology From Bed to Bench 5*, 2 (Spring 2012), 79–83.

[81] Raban, D, and Harper, F. Motivations for answering questions online. *New media and innovative technologies 73* (2008).

[82] Raina, Rajat, Ng, Andrew Y, and Koller, Daphne. Constructing informative priors using transfer learning. In *Proceedings of the 23rd international conference on Machine learning* (2006), ACM, pp. 713–720.

[83] Rao, Delip, Yarowsky, David, Shreevats, Abhishek, and Gupta, Manaswi. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents* (New York, NY, USA, 2010), SMUC '10, ACM, pp. 37–44.

[84] Ratkiewicz, Jacob, Fortunato, Santo, Flammini, Alessandro, Menczer, Filippo, and Vespignani, Alessandro. Characterizing and modeling the dynamics of online popularity. *Phys. Rev. Lett. 105* (Oct 2010), 158701.

[85] Rattigan, Matthew J., Maier, Marc E., and Jensen, David D. Relational blocking for causal discovery. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011* (2011).

[86] Reis, Virgile Landeiro Dos, and Culotta, Aron. Using matched samples to estimate the effects of exercise on mental health from twitter. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015), AAAI'15, AAAI Press, pp. 182–188.

[87] Robins, James. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling 7*, 9-12 (1986), 1393–1512.

[88] Rosenbaum, Paul R., and Rubin, Donald B. The central role of the propensity score in observational studies for causal effects. *Biometrika 70* (1983), 41–55.

[89] Rubin, Donald B, and Thomas, Neal. Matching using estimated propensity scores: Relating theory to practice. *Biometrics* (1996), 249–264.

[90] Ruths, Derek, and Pfeffer, Jürgen. Social media for large studies of behavior. *Science 346*, 6213 (2014), 1063–1064.

[91] Salganik, Matthew J., Dodds, Peter Sheridan, and Watts, Duncan J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science 311*, 5762 (2006), 854–856.

[92] Scheines, Richard, and Ramsey, Joseph. Measurement error and causal discovery. *CEUR workshop proceedings 1792* (06 2016), 1–7.

[93] Shadish, W. R., Cook, T. D., and Campbell, D. T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston, MA, 2002.

[94] Shalizi, Cosma R., and Thomas, Andrew C. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research 40* (Nov. 2010), 211–239.

[95] Sharma, Amit, Hofman, Jake M., and Watts, Duncan J. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation* (New York, NY, USA, 2015), EC '15, ACM, pp. 453–470.

[96] Silva, Ricardo, Scheine, Richard, Glymour, Clark, and Spirtes, Peter. Learning the structure of linear latent variable models. *J. Mach. Learn. Res. 7* (Dec. 2006), 191–246.

[97] Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*, 2nd ed. The MIT Press, 2000.

[98] Spirtes, Peter, Meek, Christopher, and Richardson, Thomas. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (San Francisco, CA, USA, 1995), UAI'95, Morgan Kaufmann Publishers Inc., pp. 499–506.

[99] Stock, J.H., and Watson, M.W. *Introduction to Econometrics*. Addison-Wesley series in economics. Addison Wesley, 2003.

[100] Stuart, Elizabeth A. Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics 25*, 1 (2010), 1.

[101] Sundararajan, Arun, Provost, Foster, Oestreicher-Singer, Gal, and Aral, Sinan. Research commentaryâinformation in digital, economic, and social networks. *Information Systems Research 24*, 4 (2013), 883–905.

[102] Varian, Hal R. Big data: New tricks for econometrics. *The Journal of Economic Perspectives 28*, 2 (2014), 3–27.

[103] Viswanath, Bimal, Mislove, Alan, Cha, Meeyoung, and Gummadi, Krishna P. On the evolution of user interaction in facebook. In *Proceedings of the 2Nd ACM Workshop on Online Social Networks* (New York, NY, USA, 2009), WOSN '09, ACM, pp. 37–42.

[104] Wager, S., and Athey, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *ArXiv e-prints* (Oct. 2015).

[105] Watts, Duncan J. A twenty-first century science. *Nature 445*, 7127 (02 2007), 489–489.

[106] Wilkinson, Dennis M. Strong regularities in online peer production. In *Proceedings of the 9th ACM Conference on Electronic Commerce* (New York, NY, USA, 2008), EC '08, ACM, pp. 302–309.

[107] Winship, Christopher, and Sobel, Michael. *Handbook of Data Analysis*. SAGE Publications, Ltd, 2004.

[108] Yang, Qiang, Zheng, Vincent Wenchen, Li, Bin, and Zhuo, Hankz Hankui. Transfer learning by reusing structured knowledge. *AI Magazine 32*, 2 (2011), 95–106.

[109] Yosinski, Jason, Clune, Jeff, Bengio, Yoshua, and Lipson, Hod. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3320–3328.

[110] Zamal, Faiyaz Al, Liu, Wendy, and Ruths, Derek. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM* (2012), John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, Eds., The AAAI Press.

[111] Zhang, Jiji. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell. 172*, 16-17 (2008), 1873–1896.

[112] Zhang, Jun, Ackerman, Mark S, and Adamic, Lada. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web* (2007), ACM, pp. 221–230.

[113] Zhang, Kun, Peters, Jonas, Janzing, Dominik, and Schölkopf, Bernhard. Kernel-based conditional independence test and application in causal discovery. In *UAI* (2011), Fabio Gagliardi Cozman and Avi Pfeffer, Eds., AUAI Press, pp. 804–813.