Causal Understanding of Fake News Dissemination on Social Media

Lu Cheng[1], Ruocheng Guo[1], Kai Shu[2], Huan Liu[1]
[1] Computer Science and Engineering, Arizona State University, USA
[2] Department of Computer Science, Illinois Institute of Technology, USA
{lcheng35,rguo12,huanliu}@asu.edu,kshu@iit.edu

ABSTRACT

Recent years have witnessed remarkable progress towards computational fake news detection. To mitigate its negative impact, we argue that it is critical to understand what user attributes potentially *cause* users to share fake news. The key to this causal-inference problem is to identify *confounders* – variables that cause spurious associations between treatments (e.g., user attributes) and outcome (e.g., user susceptibility). In fake news dissemination, confounders can be characterized by fake news sharing behavior that inherently relates to user attributes and online activities. Learning such user behavior is typically subject to *selection bias* in users who are susceptible to share news on social media. Drawing on causal inference theories, we first propose a principled approach to alleviating selection bias in fake news dissemination. We then consider the learned *unbiased* fake news sharing behavior as the surrogate confounder that can fully capture the causal links between user attributes and user susceptibility. We theoretically and empirically characterize the effectiveness of the proposed approach and find that it could be useful in protecting society from the perils of fake news.

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; • **Human-centered computing** → **Social media**; • **Social and professional topics** → *User characteristics.*

KEYWORDS

Fake news; User behavior; Causal inference; Social media

ACM Reference Format:
Lu Cheng[1], Ruocheng Guo[1], Kai Shu[2], Huan Liu[1]. 2021. Causal Understanding of Fake News Dissemination on Social Media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3447548.3467321

1 INTRODUCTION

Online social media ushers the world to an unprecedented time of "fake news" – false or misleading information disguised in news articles to mislead consumers [12, 40]. This has raised serious concerns, demanding novel approaches to understanding fake news dissemination. While great effort can be seen in computational fake news detection, less is known about what user attributes *cause* some users to share fake news. In contrast to the research focused on correlations between user profiles (e.g., age, gender) and fake news (e.g., [41]), this work seeks a more nuanced understanding of how user profile attributes are *causally* related to user susceptibility to share fake news[1]. The key to identifying causal user attributes with observational data is to find *confounders* – variables that cause spurious associations between treatments (user profile attributes) and outcome (user susceptibility). When left out, confounders can result in biased and inconsistent effect estimations.

But what is the main source of confounding bias in fake news dissemination? Various studies in psychology and social science have shown the strong relationships of user behavior with user characteristics and activities such as information sharing, personality traits and trust [3, 6, 48]. Consequently, characterizing user behavior has become a vital means to analyzing activities on social networking sites. Informed by this, we argue that *fake news sharing behavior*, i.e., the user-news dissemination relations characterized by a bipartite graph (see Figure 1 ①), is critical to address confounding in causal relations between user attributes and susceptibility.

Learning fake news sharing behaviour is challenging because virtually all observational social media data is subject to *selection bias* due to self-selection (e.g., users typically follow what they like) and the actions of online news platforms (e.g., these platforms only recommend news that they believe to be of interest to the users) [37]. Consequently, these biased data only partially describe how users share fake news. To alleviate the selection bias, one can leverage a technique commonly used in causal inference [18], particularly, *Inverse Propensity Scoring* (IPS) [32] that creates a pseudo-population similar to data collected from an randomized experiment. In context of fake news, propensity describes the probability of a user being exposed to a fake news piece. By connecting fake news dissemination with causal inference, we can derive an unbiased estimator for learning fake news sharing behavior under selection biases.

The main contribution of this work is three-fold. First, we address a novel and important problem that complements earlier efforts on fake news detection. In particular, we seek to answer *why* people share fake news by uncovering the causal relationships between user profiles and susceptibility. Second, we show how learning fake news sharing behavior under selection biases can be approached with propensity-weighting techniques. We design three simple and

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467321

[1] As we cannot know the exact intentions of users who spread fake news (e.g., gullible or malicious users) using only observed user engagement data, we propose a measure to approximate user susceptibility as detailed in Sec. 4.3.

effective estimations of propensity score for fake news dissemination – News-, User-News- and Neural-Network-based – to learn *unbiased embeddings* of fake news sharing behavior. Third, under the *multiple causal inference* framework with mild assumptions, we propose to use the learned embeddings of fake news sharing behavior as the confounder, drawing from findings in social science. This enables us to learn a causal model that can identify causal user attributes and estimate their effects on user susceptibility.

Our contributions are validated in an extensive empirical evaluation[2]. For the first task of modeling fake news dissemination, we show that our proposed unbiased estimators improve accuracy of predicting fake news that users are more likely to spread. By comparing the learned embeddings of fake and true news sharing behavior, we make insightful findings on the differences of the two sharing behaviors. For the second task of identifying causal attributes of susceptible users, we first show that the predictive accuracy can be improved by incorporating the unbiased embeddings of fake news sharing behavior as confounders. We then reveal multiple user attributes that are potential causes of user susceptibility. The study concludes with some critical theoretical and practical implications for researchers and policy makers.

2 RELATED WORK

2.1 Fake News Detection

Established work generally falls in two categories: content-based and propagation-based methods [51]. In content-based methods, news content is typically represented by knowledge, style, or a latent representation. Knowledge-guided methods seek to directly evaluate news authenticity by comparing its knowledge with that within a knowledge graph. Fake news detection then naturally becomes a link prediction task [39]. Limited to the completeness of knowledge graphs, further post-processing approaches for knowledge inference are often required [27]. Style features can be word-level features such as TF-IDF and/or LIWC features [7, 29]. However, style-based methods are "rarely supported by fundamental theories across disciplines" [51]. Latent-representation-based methods (e.g., [49]) have limited interpretability.

Propagation-based methods advocate the use of social context information. For instance, news cascade [8] was extended by introducing user roles (i.e., opinion leaders or normal users), stance (e.g., approval or doubt) and sentiments expressed in user posts [50]. The underlying assumption is that the overall structure of fake news cascades differs from the true ones. In early detection of fake news, news cascade was used as multivariate time series to model the propagation path of each news story [25]. Another line of research focuses on self-defined graphs such as a stance graph built on user posts [19]. Fake news is then detected by mining the stance correlations within a graph optimization framework. A more common type of graphs explores relationship among news article, publishers, users, and user posts. For instance, PageRank-like algorithm [14], tensor and matrix factorization [42].

Despite the remarkable progress in detecting fake news, comparatively fewer efforts seek to understand what user profile attributes cause users to spread fake news. Here, we provide a novel *causal understanding* by learning unbiased fake news sharing behavior. This study complements earlier works by explicitly modeling fake news dissemination with a focus on combating selection bias and discovering user attributes causally related to user susceptibility.

[2]Code is available at https://github.com/GitHubLuCheng/Causal-Understanding-of-Fake-News-Dissemination.

2.2 Propensity Scoring Methods

As one of the most important techniques in causal inference, propensity score has been applied to observational studies in various fields such as medicine, economics, and computer science. The goal of propensity scoring methods is to create a pseudo-randomized trial by reweighting samples in different treatment groups using propensity scores [4] – essentially a balancing score. One of the most classical propensity scoring methods is IPS [32], where a unit's weight is equal to the inverse of its propensity score. Among all applications, ones that are most relevant to our task are causal recommender system [4, 37] and domain adaptation [9, 43]. Conventional recommender systems are subject to selection bias. Recent studies (e.g., [23, 37]) proposed to use IPS for unbiased evaluation and learning of recommender system. For instance, user preferences (inferred through ratings or user and item covariates) were used to learn unbiased estimators from biased rating data [37].

IPS has been similarly applied to domain adaptation and covariate shift. In particular, these methods reweighed the distributions of source and target domains to adjust for their distributional differences [9, 17, 44]. For instance, to address the sample selection bias, a nonparametric method was proposed to directly produce resampling weights without distribution estimation [17]. Another interpretation of IPS is importance weighting (e.g., [44]). Under the covariate shift, standard model selection techniques do not work as desired. Methods such as importance weighted cross validation (IWCV) [44] employed IPS to alleviate misestimation due to covariate shift. More recent work (e.g., [9]) further used IPS to learn domain invariant representations.

Informed by successful prior studies, in this work, we propose to leverage IPS to learn unbiased fake news sharing behavior under selection biases. We further design three simple and effective formulations to estimate propensity score in fake news dissemination. In doing so, we seek to (1) identify the causal user attributes by conditioning on the learned fake news sharing behavior; (2) study the differences between the fake and true news sharing behavior; and (3) improve models' prediction accuracy.

3 PROBLEM STATEMENT

Let $\mathcal{U} = \{1, 2, ..., u, ..., U\}$ denote users who share fake news $C = \{1, 2, ..., i, ..., N\}$. $Y_{ui} \in \mathcal{Y}$ is a binary variable representing interactions between user u and fake news i: if u spreads i, then $Y_{ui} = 1$ else, $Y_{ui} = 0$. Note that $Y_{ui} = 0$ can be interpreted as either u is not interested in i or u did not observe i. Suppose users have m profile attributes denoted by matrix $\boldsymbol{A} = (\boldsymbol{A}_1, \boldsymbol{A}_2, ..., \boldsymbol{A}_m)$. Each user u is also associated with an outcome $B \in (0, 1]$, denoting u's susceptibility to spread fake news. We aim to identify causal user attributes and estimate the effects, which consist of two tasks:

- **Fake News Sharing Behavior Learning**. *Given the user group $\mathcal{U}$, the corpus of fake news C, the set of user-fake news interactions $\mathcal{Y}$, we aim to model the fake news dissemination process and learn fake news sharing behavior $\boldsymbol{U}$* under selection biases;

- **Causal User Attributes Identification**. *Given the user attributes* $\boldsymbol{A}$*, the fake news sharing behavior* $\boldsymbol{U}$*, and the user susceptibility* B*, this task seeks to identify user attributes that potentially cause users to spread fake news and estimate the effects.*

4 THE PROPOSED FRAMEWORK

As with other observational studies, data for studying fake news is also subject to the common selection bias. In this section, we first provide mathematical formulations of the propensity-weighting model for fake news dissemination under selection biases. We then introduce three estimations of propensity score for learning unbiased embeddings of fake news sharing behavior. Under Potential Outcome framework [32], these embeddings are then used to identify the causal relationships between user attributes and susceptibility. Figure 1 features the overview of the proposed framework.

4.1 Modeling Fake News Dissemination

We begin by building a model that characterizes fake news dissemination. The key is the "implicit" feedback we collect through natural behavior such as news reading or sharing of a user with unique profile attributes. By noting which fake news a user did and did not share in the past, we may infer fake news that a user will be interested in sharing in the future. To better formulate the process of fake news dissemination, we introduce two binary variables highly related to this process: interestingness $R_{ui} \in \{0, 1\}$ and exposure $O_{ui} \in \{0, 1\}$. $R_{ui} = 1(0)$ indicates u is interested (not interested) in i; $O_{ui} = 1$ denotes user u was exposed to fake news i and $O_{ui} = 0$, otherwise. Therefore, we assume that a user spreads fake news iff s/he is both exposed to and interested in it [20]:

$$Y_{ui} = O_{ui} \cdot R_{ui}, \tag{1}$$

$$\begin{aligned} P(Y_{ui} = 1) &= P(O_{ui} = 1) \cdot P(R_{ui} = 1), \\ &= \theta_{ui} \cdot \gamma_{ui} \quad \theta_{ui} > 0; \gamma_{ui} > 0; \forall Y_{ui} \in \mathcal{Y}, \end{aligned} \tag{2}$$

where $\theta_{ui} = P(O_{ui} = 1)$ and $\gamma_{ui} = P(R_{ui} = 1)$ parameterize the probability of exposure and interestingness, respectively. As fake news dissemination is missing-not-at-random (MNAR)[3] [24], we further assume that the probability of u spreading i is represented as the product of the exposure and interestingness parameters [36].

Suppose we have a pair of fake news (i, j) with $i \neq j$ and $\mathcal{D}_{pair} = \mathcal{U} \times C \times C$ is the set of all observed (positive) interactions (u, i) and unobserved (negative) interactions (u, j). As both the interestingness variable and exposure variable are **unobserved**, the model parameters are learned by optimizing the pairwise BPR (Bayesian Personalized Ranking) loss [31] that employs user-news interactions. In doing so, we assume that the observed user-news interactions better explain users' preferences than the unobserved ones, thereby, should be assigned higher prediction scores. We first define the ideal loss function of fake news dissemination as

$$\mathcal{L}_{ideal}(\hat{S}) = \frac{1}{|\mathcal{D}_{pair}|} \sum_{(u,i,j) \in \mathcal{D}_{pair}} \gamma_{ui}(1 - \gamma_{uj}) \ell(\hat{S}_{uij}), \tag{3}$$

where $\hat{S}_{uij}$ is the difference between the predicted scores of fake news i and j, and $\ell = -\ln(\sigma(\cdot))$ represents the local loss for the triplet (u, i, j). To this end, modeling fake news dissemination is a statistical estimation problem where we seek to estimate the ideal loss functions that returns news users are most interested in using the observed user-news interactions.

4.2 Learning Unbiased Sharing Behavior

The previously introduced model for fake news dissemination directly employs user-news interactions collected from observational studies. This leads to at least two major deficiencies: first, observational data only includes positive interactions between users and fake news whereas negative interactions are never observed. Consequently, the above fake news dissemination model cannot differentiate whether unshared fake news is uninteresting to the user or has yet to be exposed to the user; second, similar to the preferential attachment theory[4] [2] in social network science, users are preferentially to interact with news that are already prevalent and online news platforms are also more likely to recommend popular news than the tail ones. Fake news dissemination models using these partially observed interactions will learn biased embeddings of the fake news sharing behavior (or user embeddings).

To handle selection bias, we propose to leverage IPS [32, 35] to learn unbiased fake news sharing behavior based on existing positive interactions between users and fake news. To recall, propensity in fake news dissemination denotes the probability of exposing a user to a fake news piece. IPS works as a reweighting mechanism by assigning larger weights to news that is less likely to be observed. Particularly, we assume that the event of user being exposed to fake news is probabilistic, i.e., the marginal probability $\theta_{ui} = P(O_{ui} = 1)$ of observing a non-zero entry Y_{ui} for all user-fake news pairs. Formally, we define the propensity score in the fake news dissemination as follows:

Definition 1 (Propensity Score)**.** The propensity score of user u being exposed to news i is

$$\theta_{ui} = P(O_{ui} = 1) = P(Y_{ui} = 1 | R_{ui} = 1). \tag{4}$$

Eq. 4 indicates that the propensity score is the probability of u spreading i given u is interested in i. This ensures that, in principle, there could be positive interaction between every pair of (u, i). Incorporating θ_{ui} into the ideal loss function of fake news dissemination, we obtain the following unbiased estimator:

$$\hat{\mathcal{L}}_{unbiased}(\hat{S}) = \frac{1}{|\mathcal{D}_{pair}|} \sum_{(u,i,j) \in \mathcal{D}_{pair}} \frac{Y_{ui}}{\theta_{ui}} \left(1 - \frac{Y_{uj}}{\theta_{uj}}\right) \ell(\hat{S}_{uij}), \tag{5}$$

Informed by the MNAR literature [20, 36], in the following proposition, we show that this propensity-score-based estimator is unbiased w.r.t. fake news dissemination.

Proposition. The loss function in Eq. 5 is unbiased against the ideal loss of fake news dissemination in Eq. 3.

$$\mathbb{E}\left[\hat{\mathcal{L}}_{unbiased}(\hat{S})\right] = \mathcal{L}_{ideal}(\hat{S}). \tag{6}$$

The proof of this proposition can be found in Appendix A.

[3] MNAR implies the probability of an event (e.g., sharing fake news) being missing/unobserved varies for reasons that are unknown to us.

[4] Preferential attachment describes a phenomenon that the connection probability to an existing node is proportional to the degree of the target node.

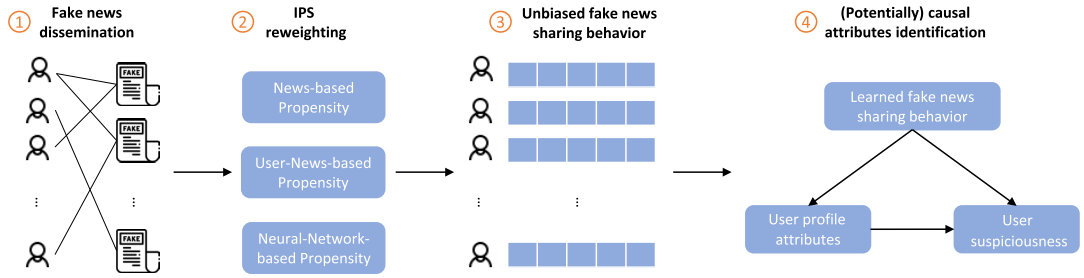


Figure 1: Overview of our framework. We model the fake news dissemination under selection biases (①) and design three effective estimations of propensity score (②) to learn unbiased embeddings of fake news sharing behavior (③). Following the causal graph with the fake news sharing behavior being the confounder (④), we examine the causal relationships between user profile attributes and susceptibility. Note that the identified attributes are "potentially" causal because as with most other observational studies, no *conclusive* causal claims can be made.

4.2.1 Propensity Score for Fake News Dissemination. Here, we propose three estimations of propensity score based on user and news attributes. The first formulation estimates propensity score using relative news popularity and is defined as

Definition 2 (News-based Propensity)**.** Propensity using relative news popularity is defined as

$$P_{news} = \hat{\theta}_{,i}^{news} = \left(\frac{\sum_{u \in \mathcal{U}} Y_{ui}}{\max_{i \in C} \sum_{u \in \mathcal{U}} Y_{ui}} \right)^{\eta}, \tag{7}$$

Typically, popularity-related measures follow power law distributions, therefore, we include the smoothing parameter $\eta \leq 1$ and set it to 0.5. With P_{news}, we assume that the probability of a user observing a fake news piece is highly related to its popularity.

Definition 3 (User-News-based Propensity)**.** Propensity using both relative news popularity and user popularity is defined as

$$P_{user} = \hat{\theta}_{u,i}^{user} = \left(\frac{\sum_{u \in \mathcal{U}} Y_{ui} \cdot F_u}{\max_{i \in C} \sum_{u \in \mathcal{U}} Y_{ui} \cdot F_u} \right)^{\eta}, \tag{8}$$

where F_u denotes the number of followers of u and $\eta = 0.5$. P_{user} also considers the bias induced by the user popularity, that is, users who are popular and active on social media are more likely to be exposed to fake news. Both estimations are input of Eq. 5. In the third formulation, we jointly estimate the propensity score and model fake news dissemination.

Definition 4 (Neural-Network-based Propensity)**.** Propensity encoded by neural networks is defined as

$$P_{neural} = \hat{\theta}_{,i}^{neural} = \sigma(\boldsymbol{e}_i), \tag{9}$$

where $\boldsymbol{e}_i$ is the latent representations of news content and $\sigma(\cdot)$ is the sigmoid function. Here, we implicitly encode the popularity of fake news in the latent space based on the news content.

4.2.2 Variance Reduction. It is widely known that IPS-based approaches often suffer from large variance as the propensity score can be extremely small. For example, fake news that is unpopular has low exposure probability. To reduce the variance, we employ the following non-negative loss [36]:

Definition 5 (Non-Negative Loss)**.** Given the propensity scores, the non-negative loss can be defined as

$$\hat{\mathcal{L}}_{non-neg}(\hat{S}) = \frac{1}{|\mathcal{D}_{pair}|} \sum_{(u,i,j) \in \mathcal{D}_{pair}} \max \left\{ \ell_{unbiased}(\hat{S}_{uij}), 0 \right\}. \tag{10}$$

Similar to the non-negative loss for Positive-Unlabeled learning with limited Positive data, Eq. 10 is more robust against the small propensity scores, and reduces the variance at the cost of introducing some bias [22]. The final loss function for modeling the unbiased fake news dissemination is formulated as follows:

$$\arg\min_{\boldsymbol{U},\boldsymbol{V}} \hat{\mathcal{L}}_{non-neg}(\hat{S}) + \lambda(\|\boldsymbol{U}\|_2^2 + \|\boldsymbol{V}\|_2^2), \tag{11}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are user embeddings (i.e., the embeddings of fake news sharing behavior) and news embeddings, respectively. λ is a hyperparameter that controls the weight of the ℓ_2-regularization for the latent factors.

4.3 Identifying Causal User Attributes

This section discusses how to simultaneously identify multiple user attributes that *potentially* cause user susceptibility and estimate the effects. Causal inference is the anchor of knowledge to understand the underlying mechanism that drives people to spread fake news [13]. With multiple user attributes at hand, we are essentially tackling a multiple causal inference task where user attributes represent the multiple treatments and user susceptibility denotes the outcome. The goal is to estimate simultaneously the effects of individual user attributes on how likely a user spread a fake news piece.

Suppose u's attributes are encoded in a vector $\boldsymbol{a} = (a_1, a_2, ..., a_m)$, $\boldsymbol{a} \in \boldsymbol{A}$. For each user u, there is a potential outcome function that maps configurations of the attributes to user susceptibility $B_u \in (0, 1]$ which is formally defined as

$$B_u = n_{fake}^u / (n_{fake}^u + n_{true}^u), \tag{12}$$

where n_{fake}^u is the number of fake news u has shared. Here we assume that *a larger portion of news a user has shared is fake, more susceptible s/he is to share fake news.*

Multiple causal inference seeks to identify the sampling distribution of the potential outcomes $B_u(\boldsymbol{a})$ for each configuration of the attributes $\boldsymbol{a}$. However, in observational studies, we can only observe one potential outcome of a user under one configuration of $\boldsymbol{a}$, a.k.a. the "fundamental problem of causal inference" [16]. Without knowing the full distribution of $B_u(\boldsymbol{a})$ for any $\boldsymbol{a}$, the inference of the outcome can be biased, i.e., $\mathbb{E}[B_u(\boldsymbol{a})] \neq \mathbb{E}[B_u(\boldsymbol{a})|\boldsymbol{A}_u = \boldsymbol{a}]$. The key is to identify *confounder* $\boldsymbol{Z}$ that simultaneously influences the causes $\boldsymbol{A}$ and the outcome B. We first introduce the following standard assumptions [32, 34] in causal inference:

Assumption (Causal Inference Assumptions)**.**

(1) The stable unit treatment value assumption (SUTVA): no interference between individuals and no different versions of a cause.
(2) The positivity or sufficient overlap assumption, that is

$$0 < p(\boldsymbol{a}_u \in \mathcal{A} | \boldsymbol{z}_u) < 1, \tag{13}$$

for all sets $\mathcal{A}$ with $p(\mathcal{A}) > 0$. It implies that given Z, the conditional probability of any vector of the causes is positive.
(3) Z is sufficiently rich to capture all variables influencing both $\boldsymbol{A}$ and B:

$$p(\boldsymbol{a}_u \in \mathcal{A} | B_u(a_1), ..., B_u(a_m), \boldsymbol{z}_u) = p(\boldsymbol{a}_u | \boldsymbol{z}_u). \tag{14}$$

When applied to identifying user attributes that cause user susceptibility, two critical questions remain to be answered:

(1) *what are the variables causally related to both the user attributes and user susceptibility?* and
(2) *In which cases, can all the three assumptions be satisfied?*

For the first question, the key is to understand the positive interactions between users and fake news, i.e., the *fake news sharing behavior* on social networking sites [45]. Decades of research in psychology and social science suggests that individual's online behavior and preferences are highly related to her personality traits [6], cultural norms [47], and her social activities such as hate propagation [33]. Therefore, drawing from these findings, we propose to use the learned unbiased embeddings of fake news sharing behavior as the surrogate confounder, that is, $\boldsymbol{U}_u \approx \boldsymbol{Z}_u$. The underlying assumption is that users' behavior of sharing fake news is sufficient to explain both user attributes and user susceptibility. The corresponding causal graph is illustrated in ④ in Figure 1.

For the second question, to satisfy SUTVA, it is required that user susceptibility is independent of other users' attributes and same value of an attribute has the same interpretation for all users. For example, whether a user spreads fake news or not should not depend on the age and gender of any other user. The positivity assumption can be interpreted as the observed attributes values vary within the counfounder $\boldsymbol{Z}$ strata, i.e., there should be adequate exposure variability of different levels of user attributes within the strata of fake news sharing behavior. The third assumption requires the learned embedding of the fake news sharing behavior to account for all confounding bias.

Given user profile attributes $\boldsymbol{A}$, confounder $\boldsymbol{U}$ and the user susceptibility B, we build the causal model to identify the causal user attributes and estimate their effects:

$$B_u = \boldsymbol{\beta}^\top \boldsymbol{a}_u + \boldsymbol{\gamma}^\top \boldsymbol{U}_u, \tag{15}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are coefficients. $\boldsymbol{\beta}$ denotes how user attributes affect individual decision to share fake news. A positive coefficient in $\boldsymbol{\beta}$ that passes statistical significance test indicates users with larger value on this attribute are more susceptible to spread fake news.

5 EMPIRICAL EVALUATION

The empirical evaluation starts with descriptions of experimental setup, including the datasets, baselines, evaluation metrics, and implementation details. In the second part, we report results of our experiments and discuss the implications.

Table 1: Dataset statistics.

Dataset	# Real	# Fake	# Total	# Users
PolitiFact	624	432	1,056	110,127
GossipCop	16,817	5,323	22,140	194,788

5.1 Experimental Setup

Data. Two benchmark datasets[5] for fake news detection are used for evaluation: PolitiFact[6] and GossipCop[7].

- *PolitiFact.* In *PolitiFact*, political news was collected from various sources and fact-checking evaluation results, i.e., fake or real, are provided by journalists and domain experts. This dataset consists of 624 real news and 432 fake news.
- *GossipCop.* In *GossipCop*, entertainment stories were collected from various media outlets. The fact-checking evaluation results came from the rating scores on the GossipCop website. Ratings range from 0 to 10 with 0 indicating fake and 10 real. Different from PolitiFact, GossipCop intends to show more fake stories due to its entertainment purpose. The dataset consists of 16,817 real stories and 5,323 fake stories.

The basic statistics of these two datasets are shown in Table 1. For each dataset, we create the training and test datasets with a 80/20 split. The training data is randomly selected from the original data (thus biased) whilst from the rest data, we create the test data such that we expose each user to each fake news as uniformly as possible (i.e., with equal probability, thus less biased). This method, which has been advocated as the most practical way to imitate randomized experiments [4, 23], can generate data with users' decisions under random exposures. The evaluation of causal models has long been a challenging task due to the lack of ground truth. By creating the distributional differences between the training and test data, we can compare a causal method with a non-causal method using the prediction accuracy across different environments. A causal method is expected to be more robust to the distribution shift as it is more transportable and domain-invariant [28, 30].

Baselines. We are not aware of any similar work in the literature of fake news that learns the embeddings of fake news sharing behavior and identifies causal user profile attributes. As our problem setting is closely related to recommender systems, here, we employ two standard approaches in recommender systems with implicit feedback as backbones of our model: Bayesian personalized ranking for matrix factorization (BPRMF) [31] and the neural collaborative filtering model (NCF) [15]. Note that for each baseline, our approach has three different variants corresponding to the three estimated propensity scores. For example, we incorporate the propensity scores defined in Eq. 7-9 into BPRMF and get three different variants of our model: BPRMF-N, BPRMF-U and BPRMF-Neu.

We adopt two standard evaluation metrics in recommender systems – Recall@K and NDCG@K. Recall@K measures of all fake news that were actually interesting to a user, how many the model predicted to be interesting in the top K fake news. It focuses on

[5]Both are available at https://github.com/KaiDMML/FakeNewsNet.
[6]https://www.politifact.com/
[7]https://www.gossipcop.com/

the ratio of interesting fake news that are not missed by the algorithm. NDCG@K measures the accuracy of the algorithm based on the ground truth interestingness and the predicted ranking of fake news among those ranked as the top K interesting fake news. Their formal definitions can be found in Appendix B. For the implementation, we used Tensorflow [1] and Statsmodel [38]. λ is set to 1e-2 for BPRMF-based models and 1e-3 for NCF-based models, the embedding dimension is 64 and the batch size is 1,024 for both. We employ the plain architecture of NCF, where the dimension of each hidden layer keeps the same. All the models are optimized by RMSProp Optimizer [46] with a learning rate of 1e-3 for BPRMF-based models and 1e-2 for NCF-based models. More implementation details can be found in Appendix B.

5.2 Evaluation on Fake News Dissemination

We first evaluate models for learning fake news sharing behavior. Specifically, we aim to answer the research questions below:

- How does the proposed model fare against standard recommendation models w.r.t. the performance of predicting fake news that users will share?
- How is the fake news sharing behavior different from the true news sharing behavior in the latent space?

With the distribution shift between the training and test data, the first question examines the efficacy of the proposed IPS-reweighting models. For the second question, we first visualize the learned embeddings of the fake and true news sharing behavior in the 2-D space. We then compute the Silhouette Coefficient of the clustering results based on these embeddings.

5.2.1 How does our approach fare against baselines in predicting fake news that users will spread? We compare Recall@K and NDCG@K of the two base models (i.e., BPRMF and NCF) to our models using the training and test data with distributional differences. We present the results averaged over 5 repetitions along with the relative improvement for both datasets in Table 2-5. The presented improvement on each dataset and for each evaluation measure is significant at 0.05 level. We begin by observing that indeed the imposed IPS reweighting confers an advantage to alleviating the selection bias in fake news dissemination, see, e.g., the results for *PolitiFact* with the base model NCF in Table 3. The improvement is most significant when K is small, e.g., $K = 20$. This indicates that our IPS-reweighting strategy is more effective when predicting fake news that is highly likely to be shared. All three IPS estimators can achieve the best performance w.r.t. Recall@K and NDCG@K with no evidence showing that one is most superior. User-News- and Neural-Network-based propensity, mostly, present better performance when predicting fake news across different environments. Estimating propensity using user popularity and news content may be more effective than using news popularity alone.

5.2.2 Comparing News Sharing Behavior. To learn the unbiased user behavior of sharing true news, we apply the same news dissemination model described in Section 4 to all true news and the associated users. We then extract embeddings of users who shared fake news and who only shared true news, and denote them as $\boldsymbol{U}_f$ and $\boldsymbol{U}_t$, respectively. We run BPRMF-N on *PolitiFact* as a working example and visualize $\boldsymbol{U}_f$ and $\boldsymbol{U}_t$ in 2-D space using t-SNE [26]. To ensure fair comparisons, we select users who only spread fake/true news and further conduct random sampling to make the number of both types of users equal (49,000 users). In addition to qualitative analysis, we further performed DBSCAN [11] clustering on $\boldsymbol{U}_f$ and $\boldsymbol{U}_t$, respectively. Then we compute the Silhouette Coefficient of the inferred clusters. Results are presented in Figure 2.

An important notion is that embeddings of fake news sharing behavior are more concentrated on a single primary cluster whilst those of true news sharing behavior are better separated into multiple and smaller clusters. This is also evidenced by the results of Silhouette Coefficient, value of which ranges from -1 to 1. A larger value denotes that a sample is further away from its neighboring clusters. The Silhouette Coefficient of true news sharing behavior is close to 1, indicating that the samples are well matched to their own clusters. We conclude that fake and true news sharing behavior are essentially different, also suggested by previous findings about fake news cascade [8]. Particularly, users who spread true news present more diverse behaviors whereas those spreading fake news have similar sharing behaviors. Our conclusion also echoes recent findings in social science and psychology [45] showing that people susceptible to spread fake news share key characteristics such as self-disclosure [10]) and social comparison [21].

Table 2: Performance comparisons w.r.t. predicting fake news to be shared using data *PolitiFact* and base model BPRMF (%). $p < 0.05$.

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	12.36	22.18	31.10	39.51
BPRMF-N	14.45$^{\uparrow 16.9\%}$	25.11$^{\uparrow 13.2\%}$	34.34$^{\uparrow 10.4\%}$	42.72$^{\uparrow 8.1\%}$
BPRMF-U	14.78$^{\uparrow 19.6\%}$	25.65$^{\uparrow 15.6\%}$	34.91$^{\uparrow 12.2\%}$	**43.63**$^{\uparrow 10.4\%}$
BPRMF-Neu	**14.90**$^{\uparrow 20.6\%}$	**25.83**$^{\uparrow 16.5\%}$	**35.13**$^{\uparrow 13.0\%}$	43.55$^{\uparrow 10.2\%}$

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	5.33	7.51	9.22	10.71
BPRMF-N	6.39$^{\uparrow 19.9\%}$	8.73$^{\uparrow 16.2\%}$	10.49$^{\uparrow 13.8\%}$	11.97$^{\uparrow 11.8\%}$
BPRMF-U	**6.54**$^{\uparrow 22.7\%}$	8.92$^{\uparrow 18.8\%}$	10.69$^{\uparrow 15.9\%}$	**12.21**$^{\uparrow 14.0\%}$
BPRMF-Neu	6.53$^{\uparrow 22.5\%}$	**8.93**$^{\uparrow 18.9\%}$	**10.71**$^{\uparrow 16.2\%}$	12.19$^{\uparrow 13.8\%}$

Table 3: Performance comparisons w.r.t. predicting fake news to be shared using data *PolitiFact* and base model NCF (%). $p < 0.05$.

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
NCF	9.59	18.45	27.30	36.33
NCF-N	**10.42**$^{\uparrow 8.7\%}$	**19.34**$^{\uparrow 4.8\%}$	28.58$^{\uparrow 4.7\%}$	37.07$^{\uparrow 2.0\%}$
NCF-U	10.29$^{\uparrow 7.3\%}$	19.29$^{\uparrow 4.6\%}$	27.34$^{\uparrow 0.1\%}$	34.87$^{\downarrow 4.0\%}$
NCF-Neu	10.20$^{\uparrow 6.4\%}$	19.11$^{\uparrow 3.6\%}$	**28.74**$^{\uparrow 5.3\%}$	**38.39**$^{\uparrow 5.7\%}$

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
NCF	3.72	5.66	7.35	8.94
NCF-N	4.13$^{\uparrow 11.2\%}$	6.09$^{\uparrow 7.6\%}$	**7.85**$^{\uparrow 6.8\%}$	9.36$^{\uparrow 4.7\%}$
NCF-U	**4.19**$^{\uparrow 12.6\%}$	**6.18**$^{\uparrow 9.2\%}$	7.75$^{\uparrow 5.4\%}$	9.10$^{\uparrow 1.8\%}$
NCF-Neu	4.04$^{\uparrow 8.6\%}$	5.99$^{\uparrow 5.8\%}$	7.82$^{\uparrow 6.4\%}$	**9.52**$^{\uparrow 6.5\%}$

Table 4: Performance comparisons w.r.t. predicting fake news to be shared using data *GossipCop* and base model BPRMF (%). $p < 0.05$.

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	13.31	16.38	18.77	20.8
BPRMF-N	14.92$^{\uparrow 12.2\%}$	17.61$^{\uparrow 7.5\%}$	19.70$^{\uparrow 5.0\%}$	21.52$^{\uparrow 3.5\%}$
BPRMF-U	14.97$^{\uparrow 12.6\%}$	17.70$^{\uparrow 8.1\%}$	19.73$^{\uparrow 5.1\%}$	21.58$^{\uparrow 3.8\%}$
BPRMF-Neu	**15.72**$^{\uparrow 18.2\%}$	**18.76**$^{\uparrow 14.5\%}$	**21.03**$^{\uparrow 12.0\%}$	**22.96**$^{\uparrow 10.4\%}$

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	10.52	11.32	11.86	12.30
BPRMF-N	12.38$^{\uparrow 17.7\%}$	13.11$^{\uparrow 15.8\%}$	13.60$^{\uparrow 14.7\%}$	13.97$^{\uparrow 13.6\%}$
BPRMF-U	12.22$^{\uparrow 16.2\%}$	12.95$^{\uparrow 14.4\%}$	13.42$^{\uparrow 13.2\%}$	13.81$^{\uparrow 12.3\%}$
BPRMF-Neu	**12.74**$^{\uparrow 21.1\%}$	**13.56**$^{\uparrow 19.8\%}$	**14.08**$^{\uparrow 18.7\%}$	**14.49**$^{\uparrow 17.8\%}$

Table 5: Performance comparisons w.r.t. predicting fake news to be shared using data *GossipCop* and base model NCF (%). $p < 0.05$.

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
NCF	5.87	8.01	9.72	11.63
NCF-N	7.59$^{\uparrow 29.3\%}$	9.50$^{\uparrow 18.6\%}$	11.22$^{\uparrow 15.4\%}$	12.74$^{\uparrow 9.5\%}$
NCF-U	**8.99**$^{\uparrow 53.2\%}$	**10.93**$^{\uparrow 36.5\%}$	**12.73**$^{\uparrow 31.0\%}$	**14.42**$^{\uparrow 24.0\%}$
NCF-Neu	8.36$^{\uparrow 42.4\%}$	10.53$^{\uparrow 31.5\%}$	12.39$^{\uparrow 27.5\%}$	13.97$^{\uparrow 20.1\%}$

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
NCF	4.41	4.97	5.37	5.77
NCF-N	5.96$^{\uparrow 35.1\%}$	6.50$^{\uparrow 30.8\%}$	6.91$^{\uparrow 28.7\%}$	7.23$^{\uparrow 25.3\%}$
NCF-U	**7.36**$^{\uparrow 66.9\%}$	**7.91**$^{\uparrow 59.2\%}$	**8.33**$^{\uparrow 55.1\%}$	**8.68**$^{\uparrow 50.4\%}$
NCF-Neu	6.53$^{\uparrow 48.1\%}$	7.14$^{\uparrow 43.7\%}$	7.57$^{\uparrow 41.0\%}$	7.91$^{\uparrow 37.1\%}$

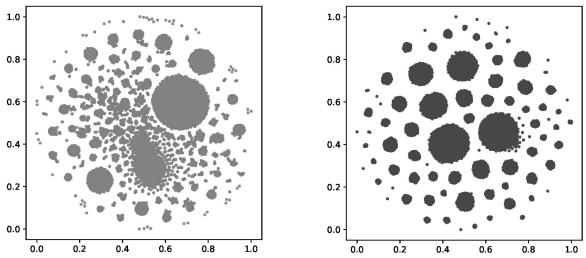
5.3 Evaluation on Identifying Causal User Attributes

We show empirical results for identifying user profile attributes that potentially cause user susceptibility to share fake news. With the unbiased embeddings of fake news sharing behavior as the confounder, in this experiment, we seek to (1) assess the effectiveness of outcome model Eq. 15 by predicting user susceptibility; meanwhile (2) discover the causal user attributes and estimate the effects. We thereby feed B_u along with $\boldsymbol{a}_u$ and $\boldsymbol{U}_u$ into Eq. 15. All the experiments in this subsection are based on the BPRMF model.

5.3.1 Effect on Predicting User Susceptibility. As the focus of this experiment is to testify the effectiveness of the unbiased fake news sharing behavior on improving predictive accuracy, here, we take the simple Linear Regression (LR)[8] as the basic model and compare the performance of LR with various input:

- *LR.* The input solely consists of the user attributes.
- *LR-Basic.* The input includes both the user attributes and embeddings of user sharing behavior learned via BPRMF.
- *LR-N.* The input includes both the user attributes and embeddings of user sharing behavior learned via BPRMF-N.

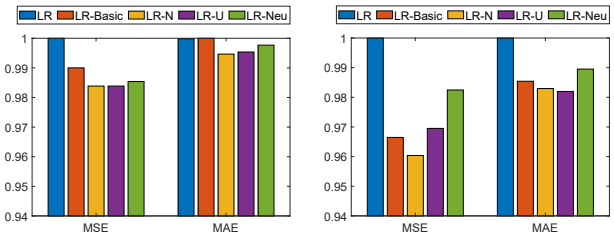
[8]This experiment can be easily adapted to other machine learning models.



(a) Fake news sharing behavior. Silhouette Coefficient=-0.124

(b) True news sharing behavior. Silhouette Coefficient=0.903

Figure 2: Behavior comparisons using 2-D t-SNE visualizations.



(a) *PolitiFact.* **(b) *GossipCop.***

Figure 3: Performance comparisons w.r.t. predicting user susceptibility using both datasets. $y-$axis denotes relative results.

- *LR-U.* The input includes both the user attributes and embeddings of user sharing behavior learned via BPRMF-U.
- *LR-Neu.* The input includes both the user attributes and embeddings of user sharing behavior learned via BPRMF-Neu.

We create the training and test data with a 80/20 split: 80% users are in the training dataset. We report the two widely used evaluation metrics for regression – Mean Squared Error (MSE) and Mean Absolute Error (MAE). The relative results are presented in Figure 3. We begin by observing that the learned embeddings of user sharing behavior can improve the accuracy of predicting user susceptibility, see, e.g., the results for *GossipCop*. Further, when taking the input of unbiased embeddings, LR can achieve the best results, especially for LR-N and LR-U. We may conclude that when predicting user susceptibility, incorporating the unbiased embeddings of fake news sharing behavior as the confounder has more positive influence on standard predictive models compared to biased embeddings.

5.3.2 Identification of Causal User Profile Attributes. In this experiment, we examine the causal relationships between user attributes and user susceptibility and estimate the effects. Specifically, we compare the coefficients of *LR-U* (i.e., debiased causal model) with those of *LR-Basic* (i.e., biased causal model) and *LR* (i.e., noncausal model). We use *LR-U* as an example because similar comparison results can be found using *LR-N* and *LR-Neu*. We present the coefficients and confidence intervals in Figures 4-5.

We first observe that *LR-U* presents more conservative estimations of the effects, see, e.g., *#status* – the number of Tweets (including retweets) issued by the user – for both datasets. This is partly because the unbiased embeddings can better alleviate the influence of confounding bias on the outcome. Additionally, *#status* has the

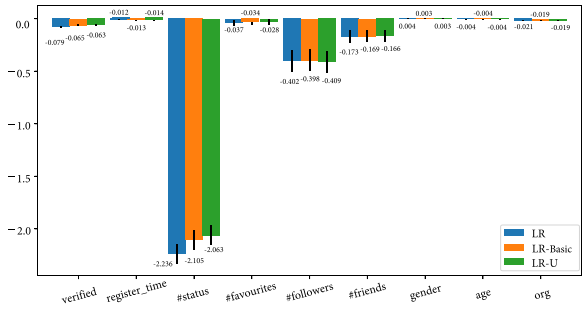


Figure 4: *PolitiFact:* Effects comparisons w.r.t. each potential causal user attribute. All the results are statistically significant.

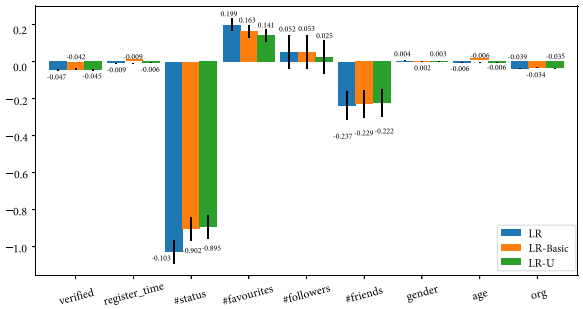


Figure 5: *GossipCop:* Effects comparisons w.r.t. each potential causal user attribute. All the results except for that of *#followers* are statistically significant.

largest effect on identifying a susceptible user, and the causal effect is negative. We may infer that users who have historically issued more tweets (regardless they are fake or not) are less susceptible to spread fake news. Similar negative effect can be observed in the binary attribute *verified* (1 denotes verified) – whether the user has a verified account, *org* (1 denotes the account represents an organization) – whether the account belongs to an organization, and *#friends* – the number of users this account is following. Intuitively, verified users and organizations are less susceptible to share fake news. While lacking ground truth for causal user attributes, by identifying profile attributes that are intuitively causes, our causal models might be applied to discovering more intrinsic user attributes that describe why people share fake news.

Our results also align well with previous findings in psychology that users with more friends share less fake news because they seek to build positive image when comparing with peers [45]. Of particular interest is that there are two contradictory results across the two datasets: effects of both *#favorites* – the number of Tweets users have liked – and *#followers* are negative in *PolitiFact* but become positive in *GossipCop*. We surmise that (1) based on the causal transpotability theory, these two attributes are less likely to be the causes of user susceptibility to share fake news. Typically, we do not need a user's approval to follow him/her on social media and following is a one-way street; (2) the category of online news platforms is a possible confounder that is left out by the surrogate confounder. While significant causal relationships are found w.r.t. *gender*, *age*, and *register_time*, the nearly zero effects warrant further studies to make causal claims regarding these attributes. This is also supported by previously conducted survey studies such as [5].

To summarize, the empirical evaluation shows that our proposed framework can learn unbiased embeddings of fake news sharing behavior that lead to more accurate predictions of fake news that users will share and user susceptibility. Our proposed causal framework also enables us to identify user attributes that potentially cause user susceptibility and estimate their effects. Comparisons between unbiased embeddings of true and fake news sharing behavior yield interesting findings regarding the differences between the two types of user behaviors.

6 DISCUSSION

We discuss the importance of understanding the causal relationships between user profile attributes and user susceptibility in combating the growing concerns about fake news. The results shown in this work demonstrate the efficacy of IPS-weighted news sharing models for learning unbiased fake news sharing behavior and the causal regression models for identifying user attributes potentially causing user susceptibility. While social media data, by itself, is not able to reliably identify the causes for why people share fake news, it can provide supporting evidence for existing conjectures and generate hypotheses for further investigation.

The observation that IPS-weighted models consistently outperform the biased fake news dissemination models in predicting fake news a user is likely to spread suggests that causal inference theories can help alleviate the selection bias to make more accurate and robust predictions. The novel results of behavioral differences between users who only spread fake news and who only spread true news enable us to develop more effective tools and techniques for detecting fake news at scale. We also study the causal relationships between user profile attributes and susceptibility to spread fake news. By incorporating unbiased embeddings of fake news sharing behavior, which can fully capture confounding between user attributes and susceptibility, the causal regression model presents better performance in predicting user susceptibility. The identified causal attributes show that *verified, statuses count, friends count, and org* relate significantly with user susceptibility to share fake news. This mirrors findings in psychology and social science and warrant future research for investigation of more intrinsic user attributes.

The results here are not without limitations: as with other studies relying on social media data, there are inherently more serious issues on selection bias that our proposed model may not be able to tackle, e.g., the selection bias in the various types of friends, differences between platforms. There is also selection bias in news that is geo-located as well as language use by the individuals on different social media platforms. It is imperative to not take these data sets as being representative of the users that may be included in the datasets. Our models are also hindered by the necessary causal inference assumptions that may be violated in practice. For instance, other unmeasured confounders (e.g., categories of social media platforms) can exist in addition to the inferred fake news sharing behavior. We do not consider other important information sources such as social networks and comments of each news. The news content and attributes have yet to be fully explored. Evaluation can

be further improved via interdisciplinary collaborations to obtain the ground-truth causal user attributes.

ETHICS STATEMENT

This work aims to advance collaborative research efforts in understanding why people spread fake news, a topic which has yet to be properly studied. Here, we provide preliminary solutions, but much work remains to bring to light the underlying causal mechanism. With our work, we hope to bring to the forefront concerns and broaden the discussions about the potential research directions in fake news. While all data used in this study are publicly available, we are committed to securing user privacy. We automatically replace user names with ordered indices in our analysis.

ACKNOWLEDGEMENTS

This material is based upon work supported by, or in part by, the U.S. Office of Naval Research and the U.S. Army Research Office under contract/grant number N00014-21-1-4002 and W911NF2020124. Kai Shu is supported by the John S. and James L. Knight Foundation through a grant to the Institute for Data, Democracy & Politics at The George Washington University.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

[3] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. 2009. Characterizing user behavior in online social networks. In *IMC*. 49–62.

[4] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. In *RecSys*. 104–112.

[5] Tom Buchanan. 2020. Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation. *Plos one* 15, 10 (2020).

[6] Laura E Buffardi and W Keith Campbell. 2008. Narcissism and social networking web sites. *Personality and social psychology bulletin* 34, 10 (2008), 1303–1314.

[7] Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A topic-agnostic approach for identifying fake news pages. In *WWW' Companion*. 975–980.

[8] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *WWW*. 675–684.

[9] Lu Cheng, Ruocheng Guo, K Selçuk Candan, and Huan Liu. 2020. Representation Learning for Imbalanced Cross-Domain Classification. In *SDM*. SIAM, 478–486.

[10] Donna Eder and Janet Lynne Enke. 1991. The structure of gossip: Opportunities and constraints on collective expression among adolescents. *ASR* (1991), 494–508.

[11] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *KDD*, Vol. 96. 226–231.

[12] Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science advances* 5, 1 (2019), eaau4586.

[13] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. 2020. A survey of learning causality with data: Problems and methods. *CSUR* 53, 4 (2020), 1–37.

[14] Manish Gupta, Peixiang Zhao, and Jiawei Han. 2012. Evaluating event credibility on twitter. In *SDM*. SIAM, 153–164.

[15] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*. 173–182.

[16] Paul W Holland. 1986. Statistics and causal inference. *JASA* 81, 396 (1986), 945–960.

[17] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. 2007. Correcting sample selection bias by unlabeled data. In *NeurIPS*. 601–608.

[18] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[19] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *AAAI*.

[20] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *WSDM*. 781–789.

[21] Susan Muller Keefer. 1994. Portrait of the gossip as a young (wo) man: Form and content of gossip among junior high school students. (1994).

[22] Ryuichi Kiryo, Gang Niu, Marthinus C Du Plessis, and Masashi Sugiyama. 2017. Positive-unlabeled learning with non-negative risk estimator. In *Neurips*. 1675–1685.

[23] Dawen Liang, Laurent Charlin, and David M Blei. 2016. Causal inference for recommendation. In *Causation: Foundation to Application, Workshop at UAI*. AUAI.

[24] Roderick JA Little and Donald B Rubin. 2019. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons.

[25] Yang Liu and Yi-Fang Brook Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *AAAI*.

[26] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR* 9, Nov (2008), 2579–2605.

[27] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2015. A review of relational machine learning for knowledge graphs. *Proc. IEEE* 104, 1 (2015), 11–33.

[28] Judea Pearl and Elias Bareinboim. 2011. Transportability of causal and statistical relations: A formal approach. In *ICDMW*. IEEE, 540–547.

[29] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104* (2017).

[30] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2015. Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint arXiv:1501.01332* (2015).

[31] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[32] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[33] Katja Rost, Lea Stahel, and Bruno S Frey. 2016. Digital social norm enforcement: Online firestorms in social media. *PLoS one* 11, 6 (2016), e0155923.

[34] Donald B Rubin. 1980. Randomization analysis of experimental data: The Fisher randomization test comment. *JASA* 75, 371 (1980), 591–593.

[35] Donald B Rubin. 2001. Using propensity scores to help design observational studies: application to the tobacco litigation. *HSORM* 2, 3-4 (2001), 169–188.

[36] Yuta Saito. 2019. Unbiased Pairwise Learning from Implicit Feedback. In *NeurIPS 2019 Workshop on Causal Machine Learning*.

[37] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. *arXiv preprint arXiv:1602.05352* (2016).

[38] Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with python. In *SciPy*, Vol. 57. Austin, TX, 61.

[39] Baoxu Shi and Tim Weninger. 2016. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-based systems* 104 (2016), 123–133.

[40] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *KDD explorations newsletter* 19, 1 (2017), 22–36.

[41] Kai Shu, Suhang Wang, and Huan Liu. 2018. Understanding user profiles on social media for fake news detection. In *MIPR*. IEEE, 430–435.

[42] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *WSDM*. 312–320.

[43] Masashi Sugiyama and Motoaki Kawanabe. 2012. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press.

[44] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. 2007. Covariate shift adaptation by importance weighted cross validation. *JMLR* 8, May (2007), 985–1005.

[45] Shalini Talwar, Amandeep Dhir, Puneet Kaur, Nida Zafar, and Melfi Alrasheedy. 2019. Why do people share fake news? Associations between the dark side of social media use and fake news sharing behavior. *JRCS* 51 (2019), 72–82.

[46] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4, 2 (2012), 26–31.

[47] Ho Keung Tsoi and Li Chen. 2011. From privacy concern to uses of social network sites: A cultural comparison via user survey. In *PASSAT*. IEEE, 457–464.

[48] Hajra Waheed, Maria Anjum, Mariam Rehman, and Amina Khawaja. 2017. Investigation of user behavior on social networking sites. *PloS one* 12, 2 (2017).

[49] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *KDD*. 849–857.

[50] Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *ICDE*. IEEE, 651–662.

[51] Xinyi Zhou and Reza Zafarani. 2018. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *arXiv preprint arXiv:1812.00315* (2018).

A PROOF OF THE PROPOSITION

Proof.

$$\begin{aligned}\mathbb{E}\big[\hat{\mathcal{L}}_{unbiased}(\hat{S})\big] &= \mathbb{E}\left[\frac{1}{|\mathcal{D}_{pair}|}\sum_{(u,i,j)\in\mathcal{D}_{pair}}\frac{Y_{u,i}}{\theta_{u,i}}\big(1-\frac{Y_{u,j}}{\theta_{u,j}}\big)\ell(\hat{S}_{uij})\right]\\ &= \frac{1}{|\mathcal{D}_{pair}|}\sum_{(u,i,j)\in\mathcal{D}_{pair}}\frac{\mathbb{E}[Y_{u,i}]}{\theta_{u,i}}\big(1-\frac{\mathbb{E}[Y_{u,j}]}{\theta_{u,j}}\big)\ell(\hat{S}_{uij})\\ &= \frac{1}{|\mathcal{D}_{pair}|}\sum_{(u,i,j)\in\mathcal{D}_{pair}}\gamma_{u,i}\big(1-\gamma_{u,j}\big)\ell(\hat{S}_{uij})\\ &= \mathcal{L}_{ideal}(\hat{S}).\end{aligned} \tag{16}$$

B REPRODUCIBILITY

B.1 Recall@K and NDCG@K in Fake News Dissemination

Recall@K is defined as

$$\text{Recall@K} = \frac{\text{\# interesting fake news @K}}{\text{Total \# interesting fake news}}. \tag{17}$$

NDCG@K is the normalized Discounted Cumulative Gain (DCG) defined as

$$\text{NDCG@K} = \frac{\text{DCG@K}}{IDCG@K}. \tag{18}$$

We define DCG@K and IDCG@K in fake news dissemination below:

$$\text{DCG@K} = \sum_{i=1}^{K}\frac{2^{interest_i}-1}{\log_2(i+1)}, \tag{19}$$

where $interest_i$ is the interestingness of the fake news piece at index i, that is, $interest_i = 1$ if a user is interested in this fake news piece and 0 otherwise. IDCG@K is the best possible value for DCG@K, i.e., the value of DCG for the best possible ranking of interesting fake news pieces at threshold K:

$$\text{IDCG@K} = \sum_{i=1}^{\substack{\text{interesting}\\ \text{news pieces at } k}}\frac{2^{interest_i}-1}{\log_2(i+1)}. \tag{20}$$

B.2 Implementation

In this section, we provide more details of the experimental setting and configuration for reproducibility purpose.

Our proposed models were implemented in Python library Tensorflow [1] and Statsmodel [38] based on two standard models for recommender system – BPRMF and NCF – as we described in Sec. 5.1. The implementation code is available at: https://github.com/GitHubLuCheng/Causal-Understanding-of-Fake-News-Dissemination. For each standard model, we have three debiased models corresponding to the three proposed propensity estimates. The file names are the combinations of the recommendation model and propensity estimates, e.g., *NCF_t.py* is the code for NCF model with news-based propensity. Implementation code for baselines is adapted from https://github.com/xiangwang1223/neural_graph_collaborative_filtering.

We used publicly available datasets for fake news, FakeNewsNet [40], available at https://github.com/KaiDMML/FakeNewsNet. For the news content, we extracted Bag of Words as features and conducted topic modeling method Latent Dirichlet Allocation (LDA) using Python package scikit-learn[9]. The extracted latent topics were then used as the input of the neural-network-based propensity estimates in Eq. 9. We detail the parameter settings for the proposed models in Table 6. The descriptions of the major parameters are introduced below:

- Emed_Size: the dimensions of user and news embeddings.
- Layer_Size: the output sizes of every layer.
- n_layers: the number of hidden layers.
- λ: the hyperparameter for ℓ_2 regularization.
- Node_Dropout: the keep probability w.r.t. node dropout for each deep layer.
- Mess_Dropout: the keep probability w.r.t. message dropout for each deep layer.
- Vocabulary_Size: the threshold to control the maximum size of vocabulary.
- n_components: the number of topics.
- max_df: when building the vocabulary ignore terms that have a document frequency strictly higher than the given threshold.
- min_df: when building the vocabulary ignore terms that have a document frequency strictly lower than the given threshold.

Table 6: Details of the parameter settings in proposed models.

Parameter	BPRMF_based model	NCF_based model
Epoch	500	500
Emed_Size	64	64
Layer_Size	64	64
Batch_Size	1024	1024
n_layers	1	1
λ	1e-2	1e-3
Learning_Rate	1e-3	1e-2
Node_Dropout	0.1	0.1
Mess_Dropout	0.1	0.1
Vocabulary_Size	2000	2000
n_components	5	5
max_df	0.5	0.5
min_df	5	5

[9] http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html