What if Social Bots Be My Friends? Estimating Causal Effect of Social Bots Using Counterfactual Graph Learning.

Ziyue Wu
Zhejiang University
ziyuewu@zju.edu.cn

Yiqun Zhang
Zhejiang University
zhangyiqun@zju.edu.cn

Xi Chen
Zhejiang University
chen_xi@zju.edu.cn

Abstract

Social bots wield significant impact within social networks. Despite the widely recognized variations in individual responses to humans and bots, existing research has not thoroughly investigated the impact differences between human and social bots on individuals' opinions. However, such differences are challenging to be estimated due to the presence of confounders introduced by homophily and the absence of counterfactual outcomes in observational network data. This study designs a counterfactual graph learning approach to accurately estimate causal effects, which exhibits superior performance in our simulations. The subsequent empirical results demonstrate that social bots yield a weaker influence than humans, and we further uncover diverse influential patterns of different types of opinions expressed by influence sources. Nevertheless, the impact difference is overestimated without applying our approach to control the confounders. Our research provides a practical approach and offers insights for stakeholders to scrutinize bots' impact from network perspectives.

Keywords: Social bots, homophily, causal identification, counterfactual graph learning.

1. Introduction

Social bots play an important role in shaping opinions within networks. These automated entities not only possess the ability to propagate ideas (Benjamin and Raghu, 2022) but can also wield the power to manipulate attitudes and emotions on social media platforms (W. Chen et al., 2021). While existing studies have explored the mechanisms by which human users influence opinion formation and change, these findings cannot be broadly extended to social bots due to their specific behavioral and social characteristics. With the expansion in network scales and the increasing complexity of online interactions, assessing the impact of these automated agents has emerged as an urgent concern for governments, platforms, and society.

Both humans and bots express opinions and exert their influence on peers through social contagion (Bapna and Umyarov, 2015; Boichak et al., 2018). However, due to inherent disparities in emotional capacities between humans and social bots (Han et al., 2022), the identical opinions expressed by these two entities may yield different impacts on adjacent human users. Specifically, when acting as influence receivers, humans can exhibit resistance to opinion voiced by AI-driven bots (Stein and Ohler, 2017) and be less susceptible to such influence in comparison to interactions with humans. This counteractive phenomena can mitigate the opinion contagion effect, resulting in a weaker influence from bots' opinions. Moreover, the variability in impact may also depend on the types of expressed opinions. In specific scenarios (such as about disasters), there may be no significant difference between humans and social bots when disseminating specific opinions (Rossi, 2022). Hence, the primary objective of this study is to assess the impact differences between social bots and humans in their capacity to shape opinions within networks.

Studies that have already looked into these impact difference have come up with inconsistent results. The variance can be attributed to the presence of homophily within social networks, wherein humans tend to establish connections with similar individuals (Bakshy et al., 2015; Kitchens et al., 2020). Homophily indicates that human users who disclose their opinions are surrounded by influence receivers who are likely

URI: https://hdl.handle.net/10125/106684
978-0-9981331-7-1

to have similar attitudes. Furthermore, these influence receivers can also be impacted by other homophilous friends. In contrast, social bots establish connections randomly or based on different rules across the whole social network (W. Chen et al., 2021; Khaund et al., 2021; Shao et al., 2018). Consequently, the observed differences in effects may indeed stem from the fact that human as the influence source are surrounded by friends with similar opinions.

To exclude the confounding effect of homophily, an effective approach is asking *what individuals' opinions would be if there exist (or not exist) bots in their neighbors* (des Mesnards et al., 2022). This counterfactual-based idea is inspired by the potential outcome framework (Rubin, 2005). In this sense, the task of accurately quantifying the impact of social bots boils down to (1) replicating the network and predicting opinions in the counterfactual scenarios, and (2) evaluating the opinion differences among users in two groups: one with bots and the other without. This idea facilitates the isolation of the distinct influence attributable to the identity of social bots.

However, solving this issue poses methodological challenges as counterfactual samples remain inaccessible in observational data. Some researchers resort to quasi-experimental strategies like matching (Bapna and Umyarov, 2015; X. Chen et al., 2022). Nevertheless, precise matching in networks is difficult due to imbalanced neighbor distributions, including neighbors' latent traits and their surrounding network structures (leading to different influences from other homophilous friends). As shown in the *Factual Graph* in Figure 1, the presence of these multiple causes contributes to the confounding bias and unidentifiable causal mechanisms.

To address these challenges, we design a novel approach called **C**ounter**f**actual **G**raph **L**earning (**CfGL**) for identifying and estimating causal effects in networks. CfGL comprises two core mechanisms: (1) Counterfactual Graph Generation, which generates counterfactual samples to simulate networks without homophily; (2) Adversarial Multi-task De-confounded Learning (AMDL) based on the factual and counterfactual graph, which address the imbalanced neighbor distributions between the treatment and control groups.

The evaluation of CfGL through simulation experiments demonstrates its superiority over existing deep graph learning methods for estimating causal effects. Then, we conduct an empirical study on Twitter network and we find that supportive opinions disclosed by humans have a greater impact on their neighbors compared to social bots, while no significant difference is observed in opposing opinions. These findings encourage researchers to further explore diverse influential patterns associated with different types of opinions. Furthermore, the application of CfGL reduces the impact difference in supportive opinion disclosure between social bots and humans, indicating that ignoring confounding factors leads to an overestimation of the differences. In summary, this study offers valuable insights to facilitate more precise estimation of bots' impact directly from observational network data. It further expands research into social bots' influence regarding outlooks within network settings.

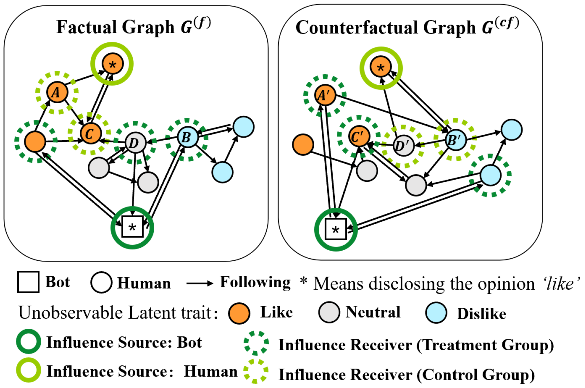
Factual Graph $G^{(f)}$
Counterfactual Graph $G^{(cf)}$
A
C
D
B
*
A'
C'
D'
B'
*
Bot
Human
Following
* Means disclosing the opinion *'like'*
Unobservable Latent trait:
Like
Neutral
Dislike
Influence Source: Bot
Influence Source: Human
Influence Receiver (Treatment Group)
Influence Receiver (Control Group)

Figure 1. An illustrative toy example. The positive opinions of A and C (compared to B and D) can be attributed to the impact differences between bots and humans, or (1) A, C have inherent positive traits (e.g., attitude); (2) A, C are affected by other homophilous friends.

2. Background and Related Work

2.1. The Impact of Social Bots

Social media platforms serve as influential channels for shaping public opinion (Ross et al., 2019). However, the precise extend of bots' influence on individuals' opinions remains debated and poorly understand. On one hand, several studies accentuate bots' significant capacity to manipulate consensus (W. Chen et al., 2021; Pescetelli et al., 2022).As public awareness of bots has grown, a majority of users believe that bots actively shape their opinions (Benjamin and Raghu, 2022). On the other hand, some research suggests that bots may not be as powerful or malevolent as initially assumed (Rossi, 2022). Regarding campaign activities on social platforms, bots often occupy less central roles versus humans, with seemingly weaker aptitude to disseminate particular stances and emotions (Cai et al., 2023).

These findings have raised concerns about bots' relative persuasiveness versus humans. Grounded in social identity theory, humans regard bots as out-groups, thus spurring the out-group bias to exhibit hostility (Castelo et al., 2019). Moreover, when social bots express subjective opinions on social media, the lay belief that bots possess diminished capacities in handling emotional and subjective issues (Castelo et al., 2019; Han et al., 2022; Longoni and Cian, 2022) reinforces this out-group bias. As a result, individuals demonstrate reluctance to embrace bots' views, instead prioritizing human opinions to preserve in-group solidarity (Castelo et al., 2019). Our study aims to precisely evaluate the distinct impacts on user opinion based on the type of influence source (i.e., human or bots).

2.2. Human and Social Bots in the Networks

Social media engagement allows both humans and social bots to select their friends. According to social network theory (SNT), human networks often exhibit ideological homophily, as humans tend to connect individuals with similar experiences and perspectives (Bakshy et al., 2015; Kitchens et al., 2020). In contrast, social bots employ various connection mechanisms based on their objectives and the specific context, such as selectively connecting with influential human users to exert their influence (Shao et al., 2018; Stella et al., 2018), or establishing random or neutral connections with humans (W. Chen et al., 2021). Bots may also interact with diverse users via assorted hashtags (Khaund et al., 2021), or just focus on certain types of individuals and contents (W. Chen et al., 2021; Salge et al., 2022). These varied mechanisms pose difficulties in establishing appropriate parametric models in advance (des Mesnards et al., 2022) to characterize bots' behavioral patterns and estimate their impacts.

The differences in forming connections between humans and bots generate distinct opinion climates (Ross et al., 2019) within their networks, posing challenges in identifying the true mechanism behind the distinctive effect mentioned earlier. Firstly, heterogeneous networks may exhibit opinion polarization due to confirmation bias (Kitchens et al., 2020), as individuals tend to accept information that aligns with their beliefs. Secondly, humans typically maintain numerous interpersonal ties beyond the association with the focal user, with viewpoints potentially bolstered via contagion among other homophilous ties (Shalizi and Thomas, 2011). In a word, this study aims to explore the impact differences between bots and humans on their neighbors' opinions in social networks by excluding the aforementioned confounders obscuring the true mechanism.

2.3. Methods for Estimating Causal Effect in Networks

Traditional approaches for causal identification and effect estimation face challenges when dealing with large-scale networks due to their limited ability to capture complex interactions and nonlinear influences (Wang et al., 2022). The advancement of neural-network-based graph learning techniques offers a promising opportunity for handling large-scale data and capturing intricate patterns in social networks (Zhou et al., 2022).

Several studies have explored the use of graph learning for estimating causal effect in networks. For instance, Veitch et al., 2019 uses network embeddings to reweights the outcome. However, the output from neural networks on the denominator will cause unstable inference. Another approach utilizes node embeddings to match the treatment and control groups (X. Chen et al., 2022). Nonetheless, achieving a precise match in both the latent traits and network structures significantly reduces the available samples. Additionally, studies suggest to balance the confounders by minimizing the representation discrepancies between two groups (Guo et al., 2020; Ma et al., 2022), which reduces the representations' ability for capturing information leading to treatment assignment bias.

Considering the needs for node representations to both align across the two groups and effectively capture information related to the treatment assignment bias, the adversarial learning emerges as a promising choice (Chu et al., 2021; Guo et al., 2021). However, these methods encounter challenges in training, as they expect the embedding vectors of each node pairs to simultaneously fulfill two opposing tasks. We solve this problem with a multi-task optimization based on counterfactual graphs and adversarial learning to de-confound the homophily bias in the networks.

3. The Proposed Framework

3.1. Problem Setting

Recall that our objective is to estimate the impact differences between social bots and humans in influencing opinions within networks. We use T to represent the treatment indicator, where $T = 1$ and $T = 0$ correspond to the disclosure of opinions by a social bot or a human, respectively. $Y^{(t)}$ is their neighbors' opinions (i.e., the potential outcome). $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{n}$

denotes observable covariates, such as demographic information. $\mathbf{U} = \{\mathbf{u}_i\}_{i=1}^n$ represents latent traits, such as inherent attitudes. A social network in the real world with n individuals can be denoted as $G^{(f)} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, ..., v_n\}$ is the node set (e.g., all the users) in the network and $e_{ij} = (v_i, v_j) \in \mathcal{E}$ is the edge (e.g., user i is followed by user j). The topology structure of the graph $G^{(f)}$ can be mathematically represented using an adjacency matrix: $\mathbf{A} = \{a_{ij}\}^{n \times n}$, where $a_{ij} \in \{0, 1\}$ and $a_{ij} = 1$ indicates the presence of an edge in the edge set $e_{ij} = (v_i, v_j) \in \mathcal{E}$.

In the context of social networks, the individual treatment effect (ITE) can be defined using the potential outcome framework (Rubin, 2005):

$$\tau_i = E[y_i^{(1)}|\mathbf{x}_i, \mathbf{u}_i, \mathbf{A}] - E[y_i^{(0)}|\mathbf{x}_i, \mathbf{u}_i, \mathbf{A}] \quad (1)$$

where $\mathbf{A} \in \{0, 1\}^{n \times n}$ is the graph adjacency matrix. τ_i reflects the impact difference between social bots and humans. To estimate ITE, it is necessary to assume that the treatment is independent of the potential outcome conditioning on $\mathbf{X}$ and $\mathbf{U}$: $\{Y_i^{(0)}, Y_i^{(1)}\} \perp T_i|\mathbf{X}, \mathbf{U}$. Since the latent homophilous factors $\mathbf{U}$ can introduce confounding bias in estimating ITE, our objective is to develop an approach that can remove the influence of $\mathbf{U}$ and obtain an accurate ITE esitmate using the observational network data $(\{\mathbf{x}_i, t_i, y_i\}_{i=1}^n, \mathbf{A})$.

3.2. Framework Overview

The review of the literatures motivates us to design a graph learning approach to identify the causal mechanism of social bots (i.e., the impact differences between humans and bots when influencing the opinions). Specifically, the Counterfactual Graph Learning (CfGL) framework aims at (1) generating counterfactual graphs and capturing homophilous factors in the network that cause the confounding bias to assist ITE estimation, and (2) balancing the neighbor distributions (nodes' latent traits and network structure) between treatment and control groups by adversarial de-confounded learning.

Figure 2 outlines the framework of CfGL. Given the input graph (i.e., *Factual Graph* $G^{(f)}$), CfGL first shuffles the edges and generate a *Counterfactual Graph* $G^{(cf)}$. Next, the **Graph Feature Extractor** processes the original *Factual Graph* $G^{(f)}$ and the synthetic *Counterfactual Graph* $G^{(cf)}$ using a graph neural network (GNN) with shared parameters and generates node embeddings $\mathbf{z}^f$ and $\mathbf{z}^{cf}$ in the same latent space. The embeddings from $G^{(f)}$ are used to predict the opinions by the **Outcome Predictor**. Additionally, the **Treatment Predictor** discriminates nodes from treatment and control groups, and the **Counterfactual Discriminator** determines whether the node comes from $G^{(f)}$ or $G^{(cf)}$. A gradient reversal layer (GRL) is placed before the Counterfactual Discriminator, which reverses the backpropagated gradient to help generate node embeddings that are indistinguishable by the discriminator. The Outcome Predictor, Treatment Predictor and Counterfactual Discriminator are trained jointly using a multi-task learning approach that can de-confound the confounding bias invoked by homophily.

3.3. Counterfactual Graph Generation

In the construction of counterfactual samples for ITE estimation, we generate counterfactual graphs $G^{(cf)}$ by shuffling n_H edges in $G^{(f)}$, with the number of shuffled edges being fewer than total edges in the original graph. This process ensures that human users are no longer selectively connected with similar individuals in the counterfactual scenario.

To extract features from both factual and counterfactual samples, we employ a representation function h that maps both $G^{(f)}$ and $G^{(cf)}$ the into the same latent space: $h : \mathcal{X} \times \mathcal{A} \rightarrow R^d$. The function h is parameterized using a Graph Attention Networks (GAT) (Veličković et al., 2017) with the following layer-wise propagation rule:

$$\mathbf{z}_i^{(l+1)} = \sigma(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^{(l)} \theta_{g_w}^{(l)} \mathbf{z}_j^{(l)}) \quad (2)$$

where θ_{g_w} is the learnable weights and σ is the activation function. $\alpha_{ij}^{(l)}$ is the attention score calculated by:

$$\alpha_{ij}^{(l)} = \frac{\exp\left(LReLU(\theta_{g_v}{}^{(l)^T}[\theta_{g_u}^{(l)} \mathbf{z}_i^{(l)}; \theta_{g_u}^{(l)} \mathbf{z}_l])\right)}{\sum_{j' \in \mathcal{N}_i} \exp\left(LReLU((\theta_{g_v}{}^{(l)^T}[\theta_{g_u}^{(l)} \mathbf{z}_i^{(l)}; \theta_{g_u}^{(l)} \mathbf{h}_{j'}^{(l)}])\right)} \quad (3)$$

where ';' denotes concatenation. We use θ_g to represent all the learnable parameters $(\theta_{g_u}, \theta_{g_v}, \theta_{g_w})$ in the GAT. The node embeddings obtained from $G^{(f)}$ and $G^{(cf)}$ serve as the outputs of the L-th layer in the GAT: $\mathbf{Z}^f, \mathbf{Z}^{cf} \in R^{N \times d}$. The adaptive learning of neighbor importance by the GAT model is useful in handling complex network structures. By ensuring that the learned embeddings of $G^{(f)}$ and $G^{(cf)}$ are in the same latent space, the model can unveil the underlying distinctions between the counterfactual and real-world network structures, which contributes to capturing the key factor that causes bias (i.e., homophily).

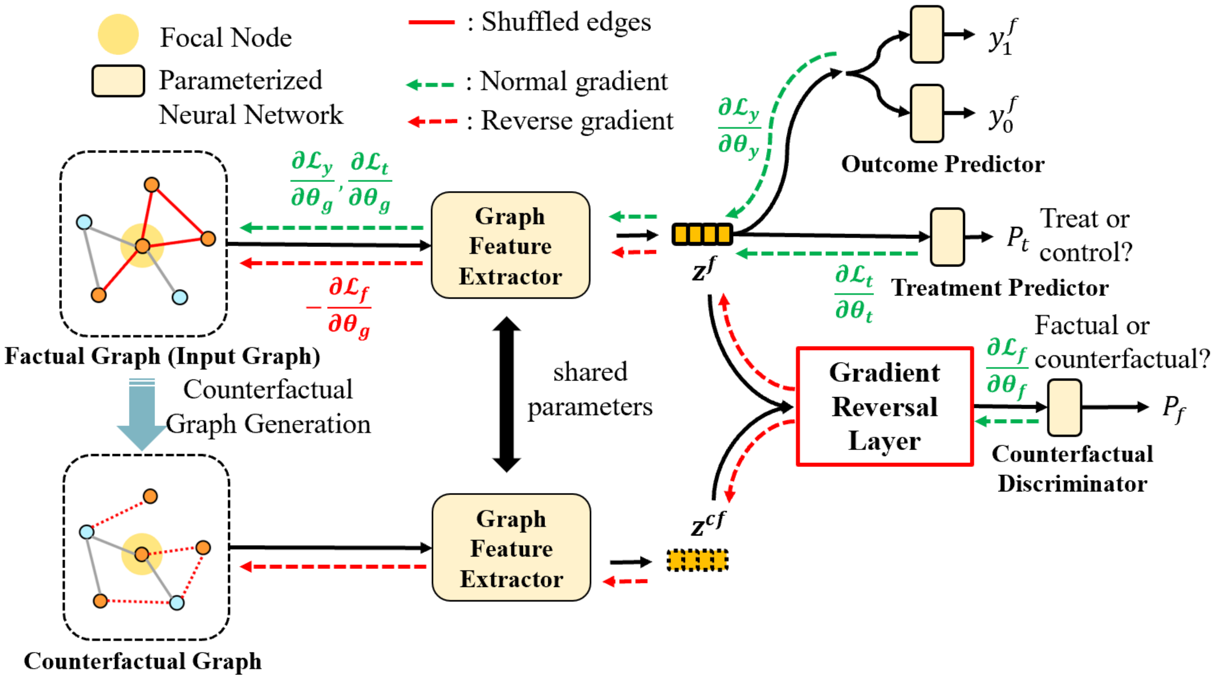
Focal Node
Parameterized Neural Network
: Shuffled edges
: Normal gradient
: Reverse gradient
$\frac{\partial \mathcal{L}_y}{\partial \theta_g}, \frac{\partial \mathcal{L}_t}{\partial \theta_g}$
$-\frac{\partial \mathcal{L}_f}{\partial \theta_g}$
Graph Feature Extractor
$\frac{\partial \mathcal{L}_y}{\partial \theta_y}$
y_1^f
y_0^f
Outcome Predictor
z^f
P_t
Treat or control?
$\frac{\partial \mathcal{L}_t}{\partial \theta_t}$
Treatment Predictor
Factual Graph (Input Graph)
Counterfactual Graph Generation
shared parameters
Gradient Reversal Layer
$\frac{\partial \mathcal{L}_f}{\partial \theta_f}$
Factual or counterfactual?
P_f
Counterfactual Discriminator
z^{cf}
Graph Feature Extractor
Counterfactual Graph

Figure 2. Framework overview.

3.4. Adversarial Multi-task De-confounded Learning

The Adversarial Multi-task De-confounded Learning (AMDL) is designed to learn the node embeddings that satisfy the following two properties: (1) The embeddings should capture enough information about the imbalanced treatment assignment; (2) The embeddings between the treatment and control groups should not exhibit large distribution discrepancy. Additionally, our joint training process is easier to reach the convergence since we set the adversarial task between different node pairs compared to other adversarial learning based methods that the task is set within the same node pairs (Chu et al., 2021; Guo et al., 2021).

3.4.1. Outcome Predictor. The first component of AMDL is the Outcome Predictor, namely, the mapping function $f : R^d \times \{0,1\} \to R$ that predicts the outcome based on node embeddings of the factual graph $G^{(f)}$:

$$\hat{y}^f = f_1(\mathbf{Z}^f), \hat{y}^f = f_0(\mathbf{Z}^f) \tag{4}$$

We parameterized f_1 and f_0 with two-layer fully-connected networks (denoted as $f_{\theta_{y1}}$ and $f_{\theta_{y0}}$) with ReLU activation function. The loss function of the outcome predictor is:

$$\mathcal{L}_y = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i^f - y_i^f)^2 \tag{5}$$

which means that the error between the predicted outcome and the true outcome is minimized during model training.

3.4.2. Treatment Predictor. To capture information that causes imbalanced treatment assignment in $G^{(f)}$, the second designed component of AMDL is the Treatment Predictor: function $f : R^d \times \{0,1\} \to P$ that maps the node embeddings to the probability of receving the treatment for each unit. The treatment predictor is trained using the following cross entropy loss:

$$\mathcal{L}_t = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{1}(p_{ti}\log(\hat{p}_{ti}) + (1-p_{ti})\log(1-\hat{p}_{ti})) \tag{6}$$

where $\hat{p}_{ti} = f_{\theta_t}(\mathbf{Z}^f)$. $p_{ti} \in \{0,1\}$ is the treatment indicator. We use a two-layer fully-connected network with parameter θ_t and LeakyReLU as the activation function.

3.4.3. Counterfactual Discriminator with a gradient reversal layer. The third component of

AMDL is the Counterfactual Discriminator with gradient reversal layers that aims at balancing the confounders between the treatment and control groups. The Counterfactual discriminator operates as a mapping function denoted as $f : R^d \times \{0,1\} \to P$, which tries to distinguish whether a node is from the factual graph $G^{(f)}$ or the counterfactual graph $G^{(cf)}$. f is parameterized by a single-layer fully-connected network with parameter θ_f, and it outputs the probability of a unit belonging to the factual graph: $\hat{p}_{fi} = f_{\theta_f}(concat[\mathbf{Z}^f; \mathbf{Z}^{cf}])$. The Counterfactual Discriminator is trained using the cross-entropy loss:

$$\mathcal{L}_f = \frac{1}{n}\sum_{i=1}^{n}\sum_{f=0}^{1}(p_{fi}\log(\hat{p}_{fi}) + (1-p_{fi})\log(1-\hat{p}_{fi})) \tag{7}$$

The GRL layer reverses the gradient from the Counterfactual Discriminator during the backpropagation. This design can helps the Graph Feature Extractor generate node embeddings that are hard to be distinguished from factual and counterfactual graph, so as to balance the confounders.

3.4.4. Joint Training. All the parameterized components (yellow-shaded module shown in Figure 2) in CfGL is jointed optimized. The final loss function of our approach can be express as:

$$\mathcal{L} = \mathcal{L}_y + \lambda_t\mathcal{L}_t + \lambda_f\mathcal{L}_f \tag{8}$$

where λ_t and λ_f are the regularized parameters. The joint training approach enables CfGL to estimate the causal effects of social bots using observational network data in an end-to-end manner.

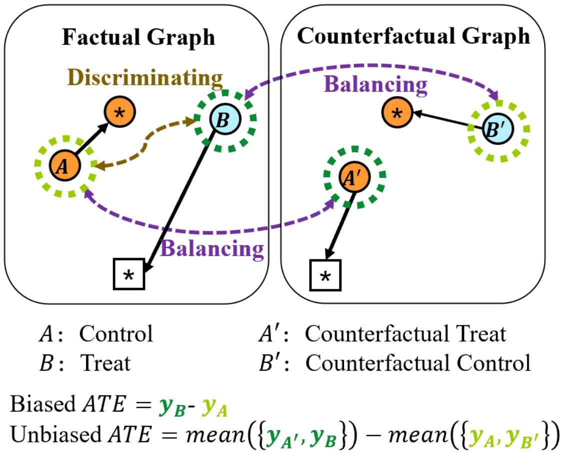

Figure 3. An illustration of the joint training approach.

We provide an illustration of the advantages of setting the adversarial task between different node pairs in CfGL in Figure 3, using nodes A and B in Figure 1 as examples. As mentioned above, the purpose of the adversarial training is to learn the node embeddings that are both discriminative (to capture information related to imbalanced treatment assignment) and indistinguishable (retaining similar between treatment and control groups). Applying adversarial training to the same node pairs, such as between nodes A and B (Guo et al., 2021), or between A and $A^{'}$ (Chu et al., 2021), will lead to difficulties in training and convergence due to the somewhat contradictory nature of the two tasks. In CfGL, the Treatment Predictor discriminates the embeddings of node A and B in factual graph (the brown arrow), whereas the GRL-attached Counterfactual Discriminator balance the embeddings of node A and $A^{'}$ as well as B and $B^{'}$ (the purple arrows). The effectiveness of the setting of CfGL will be further demonstrated through experiments.

4. Experiments

4.1. Data and Simulation

Without access to the counterfactual outcomes, obtaining ground-truth ITE is difficult since we can only see one of the possible outcomes for each sample in reality. Therefore, to harness accurate ground-truth ITE values and validate the effectiveness of our approach, we refer to the research using simulated data (X. Chen et al., 2022, Wang et al., 2022). Specifically, we set the latent trait of a human user as $U_i \in \{-1, 0, 1\}$ (e.g., the negative, neutral, and positive attitude). Our focus lies on scenarios involving positive opinion propagation, with similar considerations extended to other situations. Suppose that humans tend to form connections friends with similar traits, that is:

$$P(A_{ij} = 1) = \frac{\exp(\alpha_0 - \alpha_1|U_i - U_j|)}{1 + \exp(\alpha_0 - \alpha_1|U_i - U_j|)} \tag{9}$$

where $\alpha_1, \alpha_0 \in R$. A human with a positive attitude has a probability of disclosing his opinion. The social bots also disclose their positive opinion. Consequently, the outcome (The positive degree of the opinion) of node i can be generated through the following equation:

$$y_{i,t+1} = \beta_u U_{i,t} + \beta_h D_{i,t} + \beta_b B_{i,t} + \epsilon_{i,t} \tag{10}$$

where $\epsilon_{i,t} \sim N(0, \sigma^2)$. $D_{i,t} = 1$ indicates that node i has at least one influence source represented by a human user, while $B_{i,t} = 1$ signifies that node i is influenced by at least one social bot. Therefore, the ground-truth impact differences between social bots and humans is calculated by $\tau = \beta_b - \beta_h$. For our experiments, we set $\beta_u = \beta_b = 1, \beta_h = 2$ and $\sigma = 0.5$, ensuring that the average impact difference is equal to 1.

Bots in different domains may select the connected users by variant mechanism since the operation and algorithm of these bots are different. To verify the performance of CfGL in different domains, we generate synthetic data of the following settings: (1) **Random**. Bots randomly connect nodes in the network (W. Chen et al., 2021); (2) **RandomU**. Bots tend to randomly connect human users (Benjamin and Raghu, 2022); (3) **HighDE**. Bots tend to connect high-degree users (Stella et al., 2018); (4) **LowDE**. Bots tend to connect low-degree users (Boichak et al., 2018); (5) Bots tend to connect with influential human users to exert their impact (Shao et al., 2018; Stella et al., 2018), such as users with high betweenness centrality (**HighBC**) or high closeness centrality (**HighCC**).

4.2. Evaluation

We use two widely adopted metrics to evaluate the performance of CfGL in estimating causal effect in networks (Guo et al., 2021; Ma et al., 2022), including Mean Absolute Error (ϵ_{ATE}) and Rooted Precision in Estimation of Heterogeneous Effect ($\sqrt{\epsilon_{PEHE}}$). These two metrics are defined as follows:

$$\epsilon_{ATE} = |\frac{1}{n}\sum_{i=1}^{n}\tau_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_i| \tag{11}$$

$$\sqrt{\epsilon_{PEHE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tau_i - \hat{\tau}_i)^2} \tag{12}$$

where $\tau_i = y_i^1 - y_i^0$ is the ground-truth ITE and $\hat{\tau}_i = \hat{y}_i^1 - \hat{y}_i^0$ is the estimated ITE output by the model.

4.3. Experiment Settings

We verify the superior performance of CfGL by comparing the most commonly used graph learning methods and the most advanced approaches that conduct causal inference in the networks, including: (1) Two classical graph neural networks, **GCN** (Kipf and Welling, 2016) and **GAT** (Veličković et al., 2017), implemented with the two-hidden-layer versions; (2) Confounder balancing approaches: **CNE-** (Veitch et al., 2019), by predicting the treatments for each unit, and **ND** (Guo et al., 2020), by minimizing the distribution discrepancy between the treatment and control groups; (3) Adversarial-learning-based approach, **Ignite** (Guo et al., 2021) and **GIAL** (Chu et al., 2021).

Moreover, to verify the effectiveness of each component in CfGL, we conducted a series of ablation experiments, including: (1) **No TP**: Model that removes the Treatment Predictor (i.e., without $\mathcal{L}_t$ during the model training); (2) **No CD**: Model that removes the Counterfactual Discriminator with gradient reversal layer (i.e., without $\mathcal{L}_f$ during the model training), this is equal to removing the counterfacutal graph and adopts original input graph only.

4.4. Estimation Results

The model comparisons are presented in Table 1. We report the 5-time-average of the experiments and our CfGL outperforms existing models on the two metrics in all domains. The ablation study further demonstrates the effectiveness of each component in our proposed approach.

The experiments reveal that despite bots may establish connections with other nodes via variant mechanisms, CfGL can effectively handle the causal effect estimation in different domains. This observation underscores that our approach can identify key factors from observational network data that foster the homophily bias, and balance these factors through the generation of counterfactual samples for causal inference. It's worth noting that both GCN and GAT show the results without controlling for any confounding factors. In this context, our approach can reduce the average bias of ATE by more than 85% compared to direct modeling of observational data.

5. Empirical study

5.1. Data description

We apply our approach to real-world multi-relational Twitter network data containing social bots (Shi et al., 2023). The dataset comprises 10199 expert-annotated nodes (7451 humans and 2748 bots) and the comprehensive overview of the dataset can be found in Shi et al., 2023.

Twitter's extensive user base and real-time nature enable rapid information dissemination and facilitate public discourse. However, the platform is susceptible to the influence of social bots, accounting for approximately 20% of accounts (Rossi, 2022), which actively shape discussions and sway public opinions. This influence is especially notable within domains such as politics, health, environment, and climate, where social bots often engage with human users to propagate specific viewpoints. More specifically, the data we used is related to the stances of humans and bots on sewage discharge (Shi et al., 2023).

Table 1. Results of Benchmark Models.

Data	Random		RandomU		HighDE		LowDE		HighBC		HighCC	
Metric	eATE	ePEHE	eATE	ePEHE	eATE	ePEHE	eATE	ePEHE	eATE	ePEHE	eATE	ePEHE
GCN	0.9162	0.9355	1.0271	1.0545	0.9748	1.0294	1.0332	1.1203	1.1427	1.3813	1.0232	1.0511
GAT	0.9973	1.0145	0.9511	0.9359	0.9441	0.9836	0.9342	0.9984	0.9732	0.9923	0.9827	0.9938
CNE-	0.4933	0.5051	0.3018	0.3164	0.3246	0.3391	0.4352	0.4478	0.4601	0.4725	0.3377	0.3514
ND	0.2433	0.2590	0.2276	0.2444	0.2603	0.2763	0.2940	0.3085	0.3640	0.3774	0.2730	0.2881
Ignite	0.2358	0.4003	0.2162	0.3192	0.1775	0.2269	0.2680	0.4305	0.1620	0.2645	0.1366	0.1811
GIAL	0.1249	0.2108	0.1787	0.2245	0.1414	0.1823	0.1647	0.1943	0.2837	0.3641	0.1183	0.2188
CfGL	**0.0987**	**0.1589**	**0.1193**	**0.1683**	**0.1002**	**0.1640**	**0.0797**	**0.1674**	**0.1315**	**0.2484**	**0.0860**	**0.1451**
No TP	0.1138	0.1849	0.1374	0.2091	0.1233	0.2165	0.0946	0.1899	0.1457	0.2635	0.0984	0.1979
No CD	0.2058	0.2686	0.2910	0.3452	0.2423	0.3174	0.2890	0.3043	0.2932	0.3757	0.2692	0.3477

5.2. Model specification

Our primary objective is to compare the influential differences of social bots and humans exert over the opinions of their neighbors. A naive way to evaluate the effect of social bots involves assigning nodes to a treatment group if they have a bot neighbor with a specific opinion, and to the corresponding control group if they have a human neighbor with the same opinion. The overlapping nodes shared between these two groups are subsequently removed. Finally, the impact of social bots is quantified by measuring the opinion disparity between the two groups. However, this approach inherently introduces bias as it overlooks the latent traits of nodes and disregards the potential influence exerted by other homophilous friends.

In our CFGL approach, the treatment is defined based on whether the influence source is a bot, and the outcome is the opinions of their respective neighbors. We incorporate observable individual covariates into our input graph as node features. It is important to note that achieving precise causal identification can be challenging due to the unavailability of individual latent traits that can potentially impact both the treatment and outcome variables.

Given the absence of access to ground-truth causal effects in the empirical data, we adopt a comparative strategy (Yin and Chen, 2020) to analyze results obtained with and without our approach. The latter signifies the 'naive' approach, which does not control for confounding factors. This comparative analysis proves effective as we have demonstrated our model's performance through simulation experiments.

5.3. Results and discussion

We report the opinion differences between the treatment group (i.e., direct neighbors of a social bot that discloses its opinion) and the control group (i.e., direct neighbors of a human user that discloses his opinion). The estimated results using the Naive approach and CfGL are shown in Table 2. To ensure the robustness of the findings, we examine the friend relationship (neighbors followed by the human or bot) and the additional follower relationship (neighbors who follow the human or bot) respectively.

Table 2. Results on Empirical data.

	Against		Support	
Diff	Naive	CfGL	Naive	CfGL
Friends	0.173	0.018	-0.757	-0.262
	(0.637)	(0.481)	(0.000)	(0.020)
Friends +	0.184	0.023	-0.926	-0.518
Followers	(0.448)	(0.357)	(0.000)	(0.015)

There are several critical findings. First of all, the influence of bots on individuals' opinions is weaker than that of humans ($difference = -0.262, p = 0.020; difference = -0.518, p = 0.015$). This observation aligns with the notion that individuals tend to place more reliance on opinions expressed by in-group humans rather than out-group bots. Moreover, our proposed method reveals that the impact differences between bots and humans are overestimated when confounding factors such as homophily and influence from homophilous friends are not adequately considered.

Secondly, influential patterns vary based on opinion types. Specifically, bots exert a significantly weaker impact when expressing supportive opinions, while no discernible difference is observed for oppositional opinions. Given the potential influence of social norms on shaping opinions (Bicchieri and Mercier, 2014), we undertake a comprehensive exploration of the contextual implications behind each opinion within our dataset. Notably, only 29.5% of humans holding a supportive opinion towards wastewater discharge, it becomes apparent that the prevailing social norm is to oppose the discharge proposal. This finding demonstrates that when bots' opinions diverge from social norms, the out-group bias is enhanced and humans are highly reluctant to accept the bots' opinions. Our finding highlights the moderating role of alignment between bots' opinions and social norms in shaping individual opinions, which means that researchers are advised to distinguish between opinion types when examining the influential differences between humans and social bots.

6. Conclusions

Evaluating the impact differences between social bots and humans on the opinions in social network is important in many scenarios. In this study, we propose CfGL based on counterfactual graph learning to solve this challenging problem. CfGL shows excellent performance on the synthetic data representing various domains. We also find that humans are more influential than bots on their neighbors' opinion. However, the difference is overestimated without controlling the confounders or making distinction between the types of opinions.

Despite these advantages and insights, there are some limitations to this study. Since there is no ground-truth causal effect, we can only leverage simulated data to prove the effectiveness of our approach. The simulation contains some common cases, and future studies can exam other cases to extend our results. Moreover, our approach is based on the homophily assumption in social networks, whereas individuals may also exhibit variety seeking behaviors that lead them to connect with dissimilar friends. Future research can explore how to assess the social bots' impact in heterophily networks. Furthermore, this work can be expanded by analyzing the confounding influence of platform content recommendation systems (Pescetelli et al., 2022) on the evaluation of social bot impact. Finally, researchers can further explore the diverse influential patterns of different opinions in social networks.

References

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, *348*(6239), 1130–1132.

Bapna, R., & Umyarov, A. (2015). Do your online friends make you pay? a randomized field experiment on peer influence in online social networks. *Management Science*, *61*(8), 1902–1920.

Benjamin, V., & Raghu, T. (2022). Augmenting social bot detection with crowd-generated labels. *Information Systems Research*.

Bicchieri, C., & Mercier, H. (2014). Norms and beliefs: How change occurs. *The complexity of social norms*, 37–54.

Boichak, O., Jackson, S., Hemsley, J., & Tanupabrungsun, S. (2018). Automated diffusion? bots and their influence during the 2016 us presidential election. *Transforming Digital Worlds: 13th International Conference, iConference 2018, Sheffield, UK, March 25-28, 2018, Proceedings 13*, 17–26.

Cai, M., Luo, H., Meng, X., Cui, Y., & Wang, W. (2023). Network distribution and sentiment interaction: Information diffusion mechanisms between social bots and human users on social media. *Information Processing & Management*, *60*(2), 103197.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Chen, W., Pacheco, D., Yang, K.-C., & Menczer, F. (2021). Neutral bots probe political bias on social media. *Nature communications*, *12*(1), 5580.

Chen, X., Liu, Y., & Zhang, C. (2022). Distinguishing homophily from peer influence through network representation learning. *INFORMS Journal on Computing*, *34*(4), 1958–1969.

Chu, Z., Rathbun, S. L., & Li, S. (2021). Graph infomax adversarial learning for treatment effect estimation with networked observational data. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 176–184.

des Mesnards, N. G., Hunter, D. S., el Hjouji, Z., & Zaman, T. (2022). Detecting bots and assessing their impact in social networks. *Operations Research*, *70*(1), 1–22.

Guo, R., Li, J., Li, Y., Candan, K. S., Raglin, A., & Liu, H. (2021). Ignite: A minimax game toward learning individual treatment effects from networked observational data. *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 4534–4540.

Guo, R., Li, J., & Liu, H. (2020). Learning individual causal effects from networked observational data. *Proceedings of the 13th International Conference on Web Search and Data Mining*, 232–240.

Han, E., Yin, D., & Zhang, H. (2022). Bots with feelings: Should ai agents express positive emotion in customer service? *Information Systems Research*.

Khaund, T., Kirdemir, B., Agarwal, N., Liu, H., & Morstatter, F. (2021). Social bots and their coordination during online campaigns: A survey. *IEEE Transactions on Computational Social Systems*, *9*(2), 530–545.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kitchens, B., Johnson, S. L., & Gray, P. (2020). Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption. *MIS quarterly*, *44*(4).

Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. *Journal of Marketing*, *86*(1), 91–108.

Ma, J., Wan, M., Yang, L., Li, J., Hecht, B., & Teevan, J. (2022). Learning causal effects on hypergraphs. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1202–1212.

Pescetelli, N., Barkoczi, D., & Cebrian, M. (2022). Bots influence opinion dynamics without direct human-bot interaction: The mediating role of recommender systems. *Applied Network Science*, *7*(1), 1–19.

Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? an agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, *28*(4), 394–412.

Rossi, S. (2022). The scamdemic conspiracy theory and twitter's failure to moderate covid-19 misinformation. *HICSS*, 1–10.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

Salge, C. A. d. L., Karahanna, E., & Thatcher, J. B. (2022). Algorithmic processes of social alertness and social transmission: How bots disseminate information on twitter. *MIS Quarterly*, *46*(1).

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research*, *40*(2), 211–239.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, *9*(1), 1–9.

Shi, S., Qiao, K., Chen, J., Yang, S., Yang, J., Song, B., Wang, L., & Yan, B. (2023). Mgtab: A multi-relational graph-based twitter account detection benchmark. *arXiv preprint arXiv:2301.01123*.

Stein, J.-P., & Ohler, P. (2017). Venturing into the uncanny valley of mind—the influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition*, *160*, 43–50.

Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, *115*(49), 12435–12440.

Veitch, V., Wang, Y., & Blei, D. (2019). Using embeddings to correct for unobserved confounding in networks. *Advances in Neural Information Processing Systems*, *32*.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, G., Li, J., & Hopp, W. J. (2022). An instrumental variable forest approach for detecting heterogeneous treatment effects in observational studies. *Management Science*, *68*(5), 3399–3418.

Yin, G., & Chen, J. (2020). Improving causal inference with text as data in empirical is research: A machine learning approach. *The 48th International Conference on Information Systems*.

Zhou, J., Liu, L., Wei, W., & Fan, J. (2022). Network representation learning: From preprocessing, feature extraction to node embedding. *ACM Computing Surveys (CSUR)*, *55*(2), 1–35.