Causal Understanding of Fake News Dissemination on Social Media

Speaker: Zi-Xin, Chen

Advisor: Jia-Ling, Koh

Date: 2021/12/14

Introduction

Motivation

- False or misleading information disguised in news articles to mislead consumers has rasied serious concerns, demanding novel approaches to understanding fake news dissemination
- Great effort can be seen in computational fake news detection, but less is known about what user attributes ***cause*** some users to share fake news

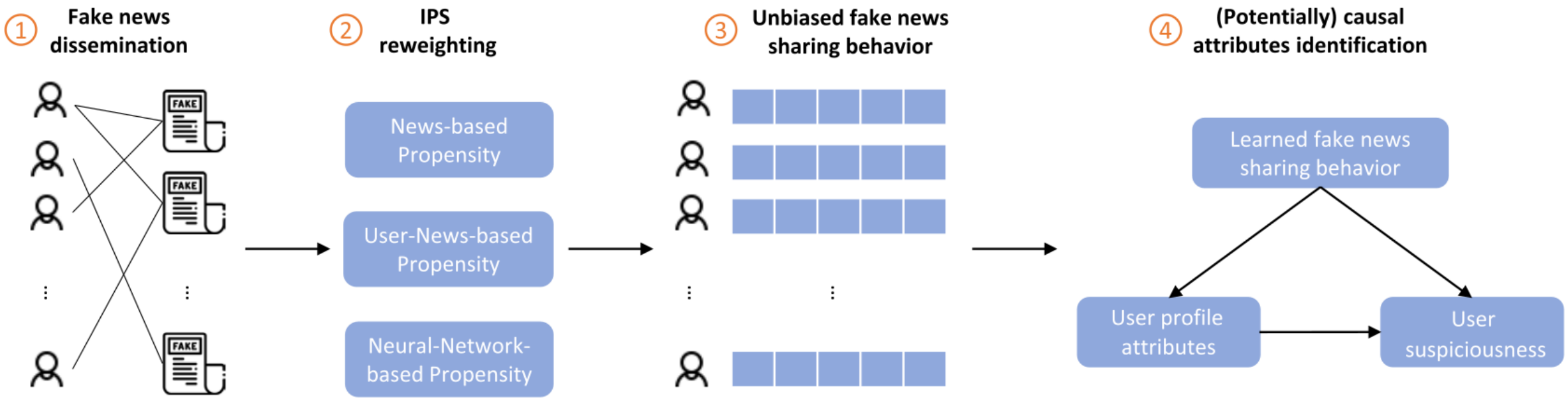
Goal

1. Propose a principled approach to alleviationg **selection bias** in fake news dissemination.
2. Then consider the leanred **unbiased** fake news sharing behavior as the surrogate confounder that can fully capture the causal links between user atttributes and user susceptibility.
3. Understand what user attributes potentially ***cause*** users to share fake news

Input and Output

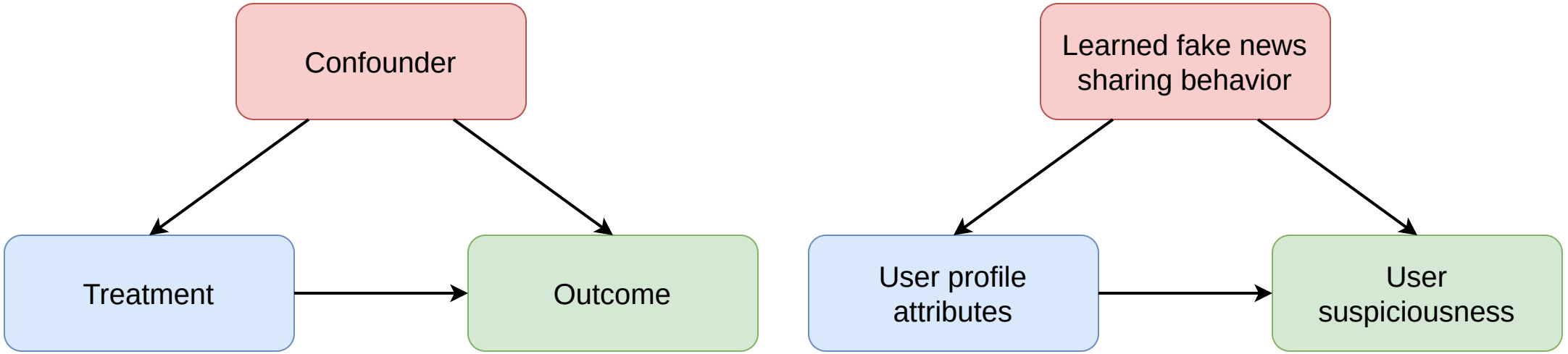
- Input: Fake news sharing behavior
- Output: User attributes that cause the sharing behavior

Overview Framework



Causal Inference

Confounder: Variables that cause spurious associations betwwen treatments and outcome



Method

Problem Statement

Users

$$\mathcal{U} = \{1, 2, ..., u, ...U\}$$

Fake News

$$\mathcal{C} = \{1, 2, ..., i..., N\}$$

Problem Statement

Interactions between user $\boldsymbol{u}$ and fake news $\boldsymbol{i}$

$$Y_{ui} \in \mathcal{Y}$$

$$Y_{ui} = \begin{cases} 1, & \text{if } \boldsymbol{u} \text{ spreads } \boldsymbol{i} \\ 0, & \text{else} \end{cases}$$

$Y_{ui} = 0$ can be interpreted as either $\boldsymbol{u}$ is note interested in $\boldsymbol{i}$ or $\boldsymbol{u}$ did not observe $\boldsymbol{i}$.

Problem Statement

Users have $\boldsymbol{m}$ profile attributes denoted by

$$\text{matrix } A = (A_1, A_2, ..., A_m).$$

Susceptibility to spread fake news of user $\boldsymbol{u}$

$$B \in (0, 1]$$

Each user $\boldsymbol{u}$ is associated with an outcome B

Problem Statement

Tasks

- Fake News Sharing Behavior Learning
  - Model the fake news dissemination process
  - Learn fake news sharing behavior U under selection biases
- Causal User Attributes Identification
  - Identify user attributes that potentially cause users to spread fake news and estimate the effects.

Modeling Fake News Dissemination

Interestingness

$$R_{ui} \in \{0, 1\}$$

Exposure

$$O_{ui} \in \{0, 1\}$$

Modeling Fake News Dissemination

Assume that a user spreads fake news iff s/he is both exposed to and interested in it

$$Y_{ui} = O_{ui} \cdot R_{ui}$$

$$P(Y_{ui} = 1) = P(O_{ui} = 1) \cdot P(R_{ur} = 1) = \theta_{ui} \cdot r_{ui}, \theta_{ui} > 0; r_{ui} > 0; \forall Y_{ui} \in \mathcal{Y}$$

$$\mathcal{D}_{pair} = \mathcal{U} \times \mathcal{C} \times \mathcal{C}$$

is the set of all observed(positive) interactions (u, i) and all unobserved(negative) interactions (u, j)

Modeling Fake News Dissemination

Optimizing the pairwise BPR loss

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

where $\hat{S}_{ui}$ 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)

Learning Unbiased Sharing Behavior

Inverse Probability Weighting (IPS)

Reweighting machanism by assigning larger weights to news that is **less** likely to be observed

Propensity Score

The propensity score of user $\boldsymbol{u}$ being exposed to news $\boldsymbol{i}$ is

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

Learning Unbiased Sharing Behavior

Unbiased estimator

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

Proposition

$$\mathbb{E}[\hat{\mathcal{L}}_{unbiased}(\hat{S})] = \mathcal{L}_{ideal}(\hat{S})$$

News-based Propensity

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

User-News-based Propensity

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

Neural-Network-based Propensity

$$P_{\text{neural}} = \hat{\theta}^{\text{neural}}_{,i} = \sigma(e_i).$$

Identifying Causal User Attributes

User $\boldsymbol{u}$'s attribute

$$a = (a_1, a_2, ..., a_m), a \in A$$

User susceptibility

$$B_u \in (0, 1]$$

Identifying Causal User Attributes

Assume a large portion of news a user has shared is fake, more ***susceptible*** s/he is to share fake news.

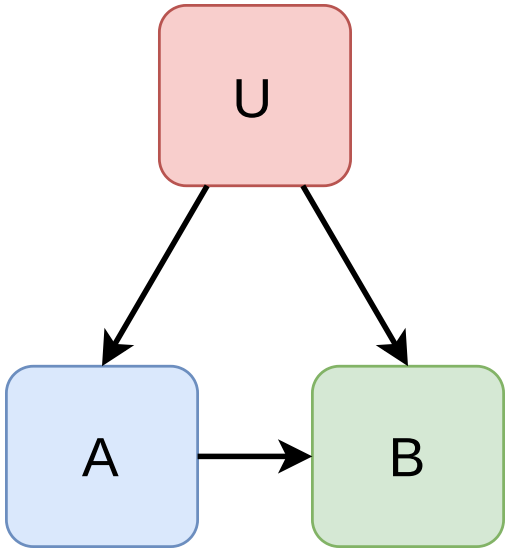
$$B_u = n^u_{\text{fake}} / (n^u_{\text{fake}} + n^u_{\text{true}})$$

$$\text{where } n^u_{\text{fake}} \text{ is the number of fake news } u \text{ has shared.}$$

Identifying Causal User Attributes

Causal Model

$$B_u = \beta^\intercal a_u + \gamma^\intercal U_u$$



Experiment

Dataset

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

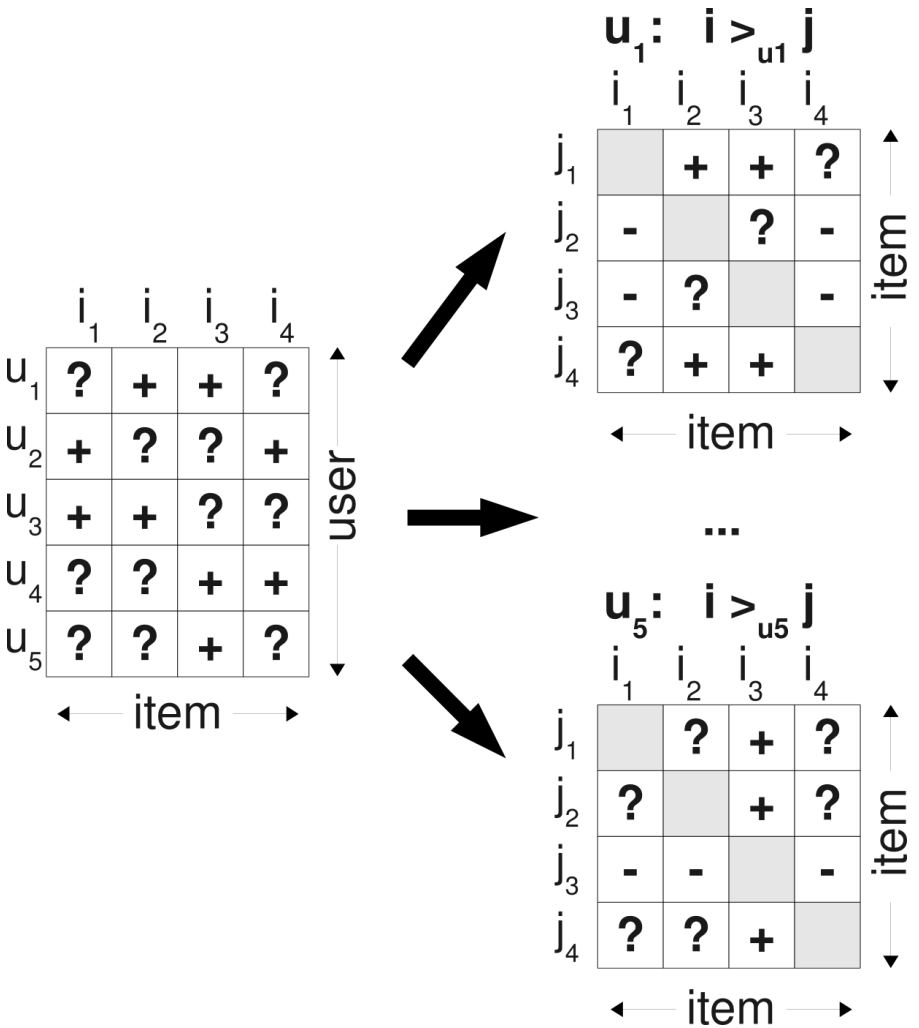
- PolitiFact: Fake or real, are provided by journalists and domain experts.
- GossipCop: The fact-checking evaluation results came from the rating scores on the GossipCop website.

Baseline

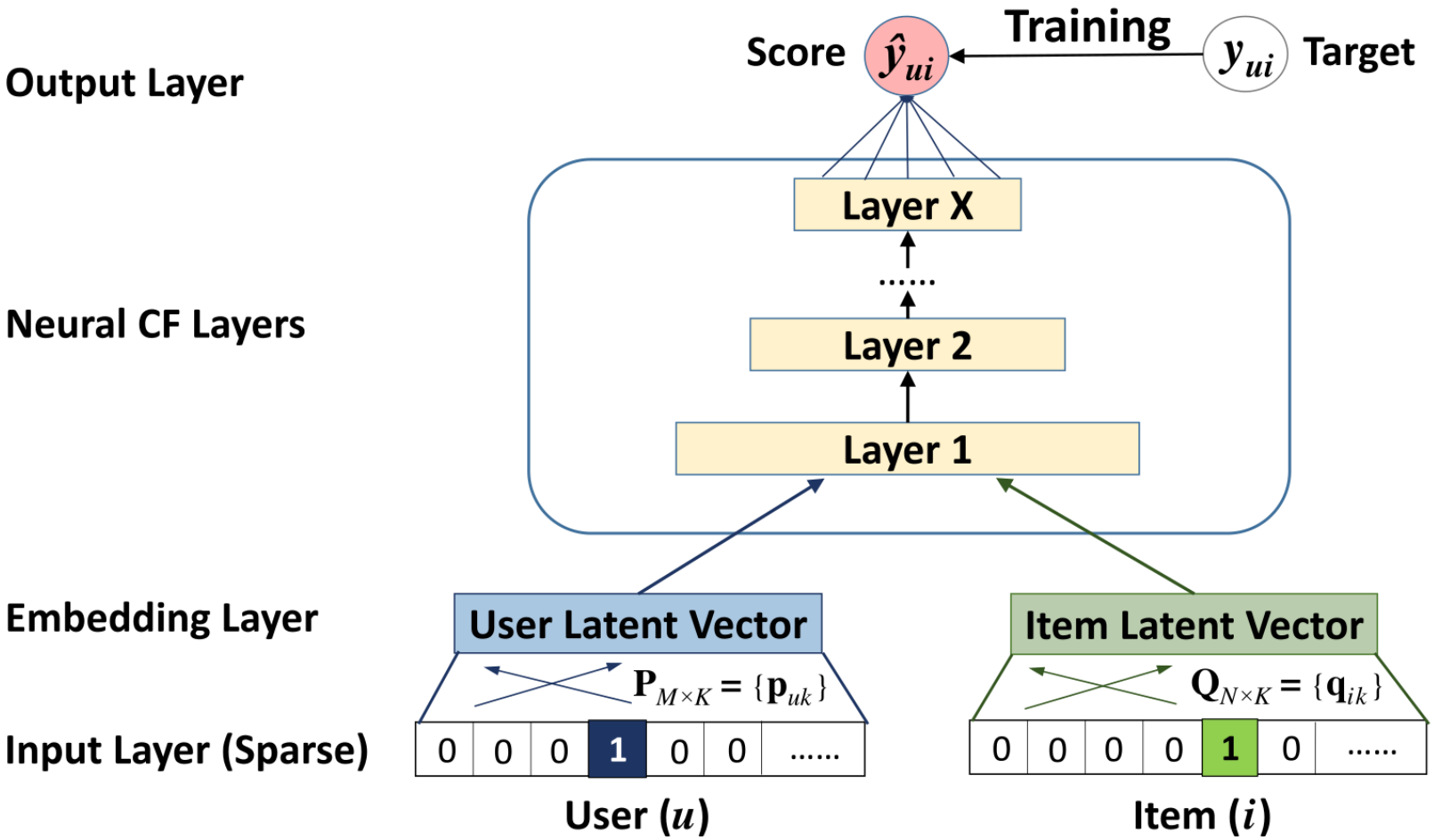
Closely related to recommendder system

- **BPRMF** (Bayesian Personalized Ranking for Matrix Factorization)
- **NCF** (Neural Collaborative Filtering Model)

BPRMF



NCF


Output Layer
Score
$\hat{y}_{ui}$
Training
y_{ui}
Target
Neural CF Layers
Layer X
.....
Layer 2
Layer 1
Embedding Layer
User Latent Vector
$\mathbf{P}_{M\times K} = \{\mathbf{p}_{uk}\}$
Item Latent Vector
$\mathbf{Q}_{N\times K} = \{\mathbf{q}_{ik}\}$
Input Layer (Sparse)
0 0 0 1 0 0 .....
0 0 0 0 1 0 .....
User (u)
Item (i)

Evaluation Metrics

- Recall@K

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

- NDCG@K

$$\text{NDCG@K} = \frac{\text{DCG@K}}{\text{IDCG@K}}, \text{DCG@K} = \sum_{i=1}^{K} \frac{2^{\text{interest}_i} - 1}{\log_2(i+1)}$$

Evaluation on Fake News Dissemination

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	$14.78^{\uparrow 19.6\%}$	$25.65^{\uparrow 15.6\%}$	$34.91^{\uparrow 12.2\%}$	$\mathbf{43.63}^{\uparrow 10.4\%}$
BPRMF-Neu	$\mathbf{14.90}^{\uparrow 20.6\%}$	$\mathbf{25.83}^{\uparrow 16.5\%}$	$\mathbf{35.13}^{\uparrow 13.0\%}$	$43.55^{\uparrow 10.2\%}$

(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	$\mathbf{6.54}^{\uparrow 22.7\%}$	$8.92^{\uparrow 18.8\%}$	$10.69^{\uparrow 15.9\%}$	$\mathbf{12.21}^{\uparrow 14.0\%}$
BPRMF-Neu	$6.53^{\uparrow 22.5\%}$	$\mathbf{8.93}^{\uparrow 18.9\%}$	$\mathbf{10.71}^{\uparrow 16.2\%}$	$12.19^{\uparrow 13.8\%}$

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	$\mathbf{15.72}^{\uparrow 18.2\%}$	$\mathbf{18.76}^{\uparrow 14.5\%}$	$\mathbf{21.03}^{\uparrow 12.0\%}$	$\mathbf{22.96}^{\uparrow 10.4\%}$

(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	$\mathbf{12.74}^{\uparrow 21.1\%}$	$\mathbf{13.56}^{\uparrow 19.8\%}$	$\mathbf{14.08}^{\uparrow 18.7\%}$	$\mathbf{14.49}^{\uparrow 17.8\%}$

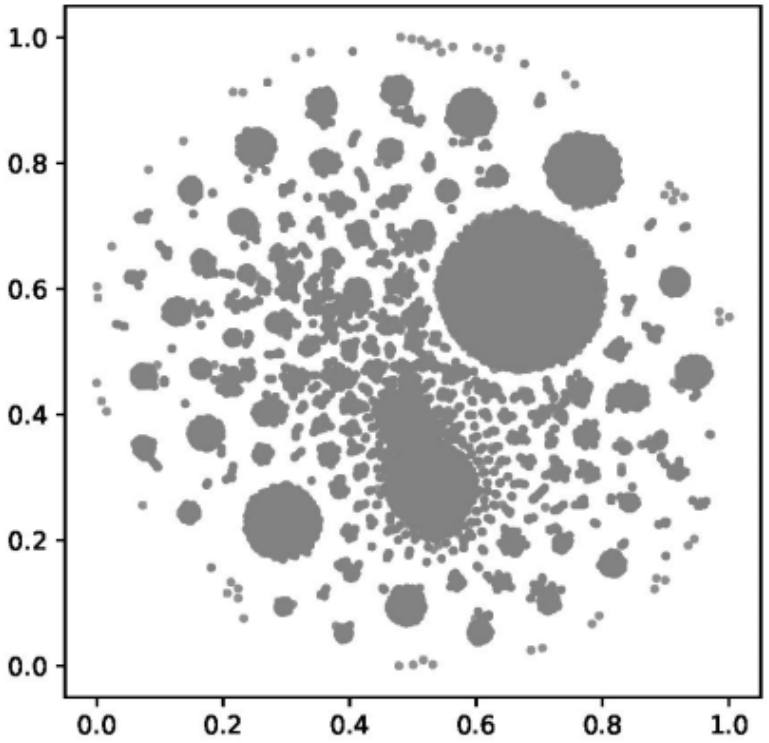
Table 5: Performance comparisons w.r.t. predicting fake news to be shared using data *GossipCop* and base model NCF (%). $p < 0.05$.

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

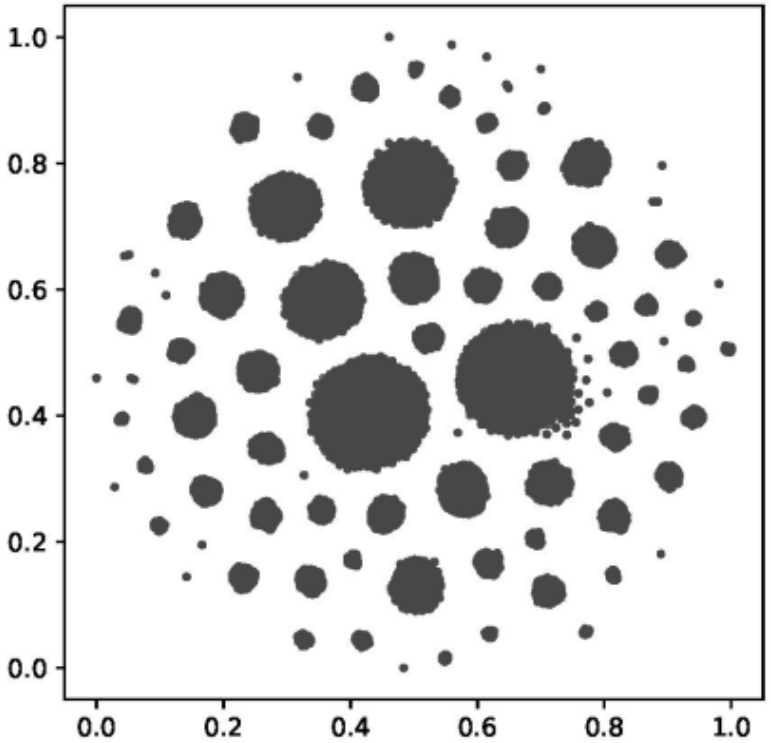
K	20	40	60	80
NCF	5.87	8.01	9.72	11.63
NCF-N	$7.59^{\uparrow 29.3\%}$	$9.50^{\uparrow 18.6\%}$	$11.22^{\uparrow 15.4\%}$	$12.74^{\uparrow 9.5\%}$
NCF-U	$\mathbf{8.99}^{\uparrow 53.2\%}$	$\mathbf{10.93}^{\uparrow 36.5\%}$	$\mathbf{12.73}^{\uparrow 31.0\%}$	$\mathbf{14.42}^{\uparrow 24.0\%}$
NCF-Neu	$8.36^{\uparrow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

(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	$\mathbf{7.36}^{\uparrow 66.9\%}$	$\mathbf{7.91}^{\uparrow 59.2\%}$	$\mathbf{8.33}^{\uparrow 55.1\%}$	$\mathbf{8.68}^{\uparrow 50.4\%}$
NCF-Neu	$6.53^{\uparrow 48.1\%}$	$7.14^{\uparrow 43.7\%}$	$7.57^{\uparrow 41.0\%}$	$7.91^{\uparrow 37.1\%}$



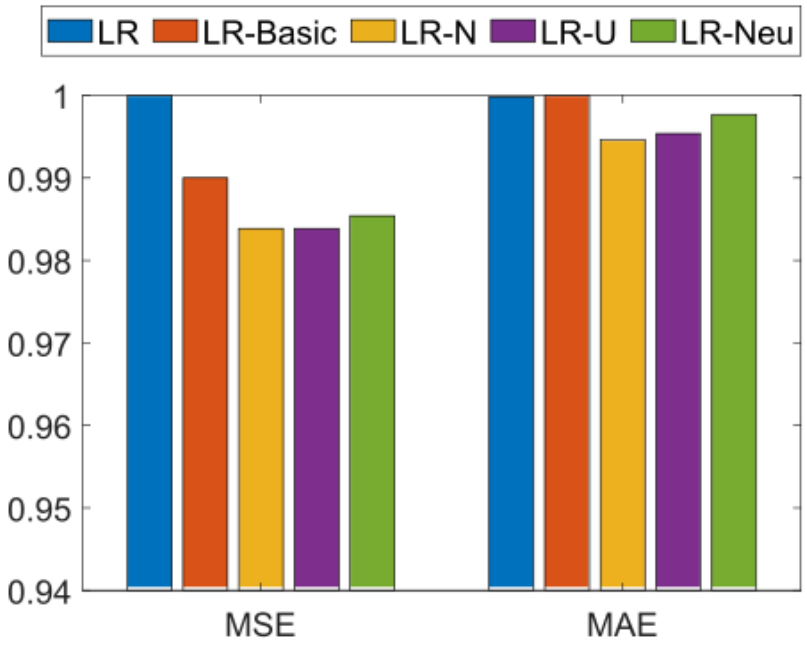
(a) Fake news sharing behavior.
Silhouette Coefficient=-0.124



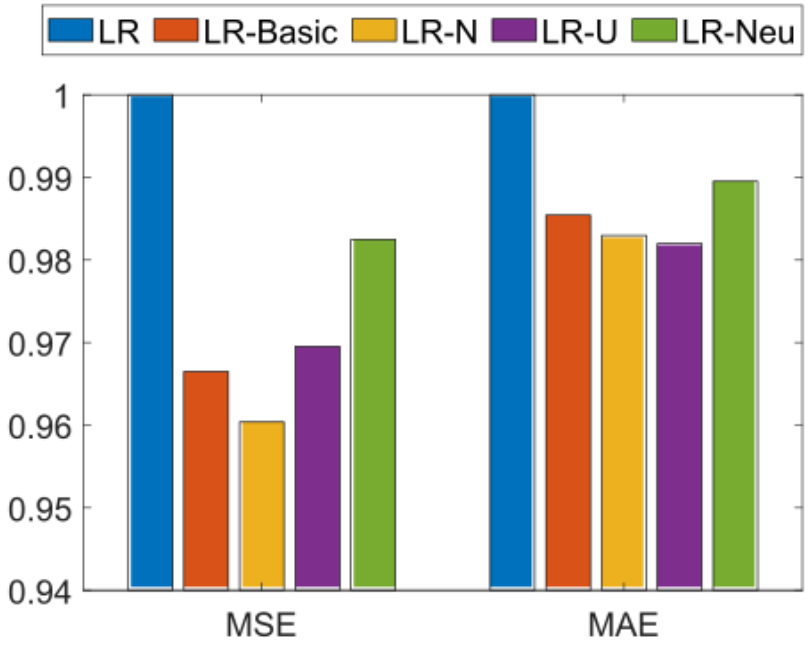
(b) True news sharing behavior.
Silhouette Coefficient=0.903

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

Evaluation on Identifying Causal User



(a) *PolitiFact.*



(b) *GossipCop.*

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

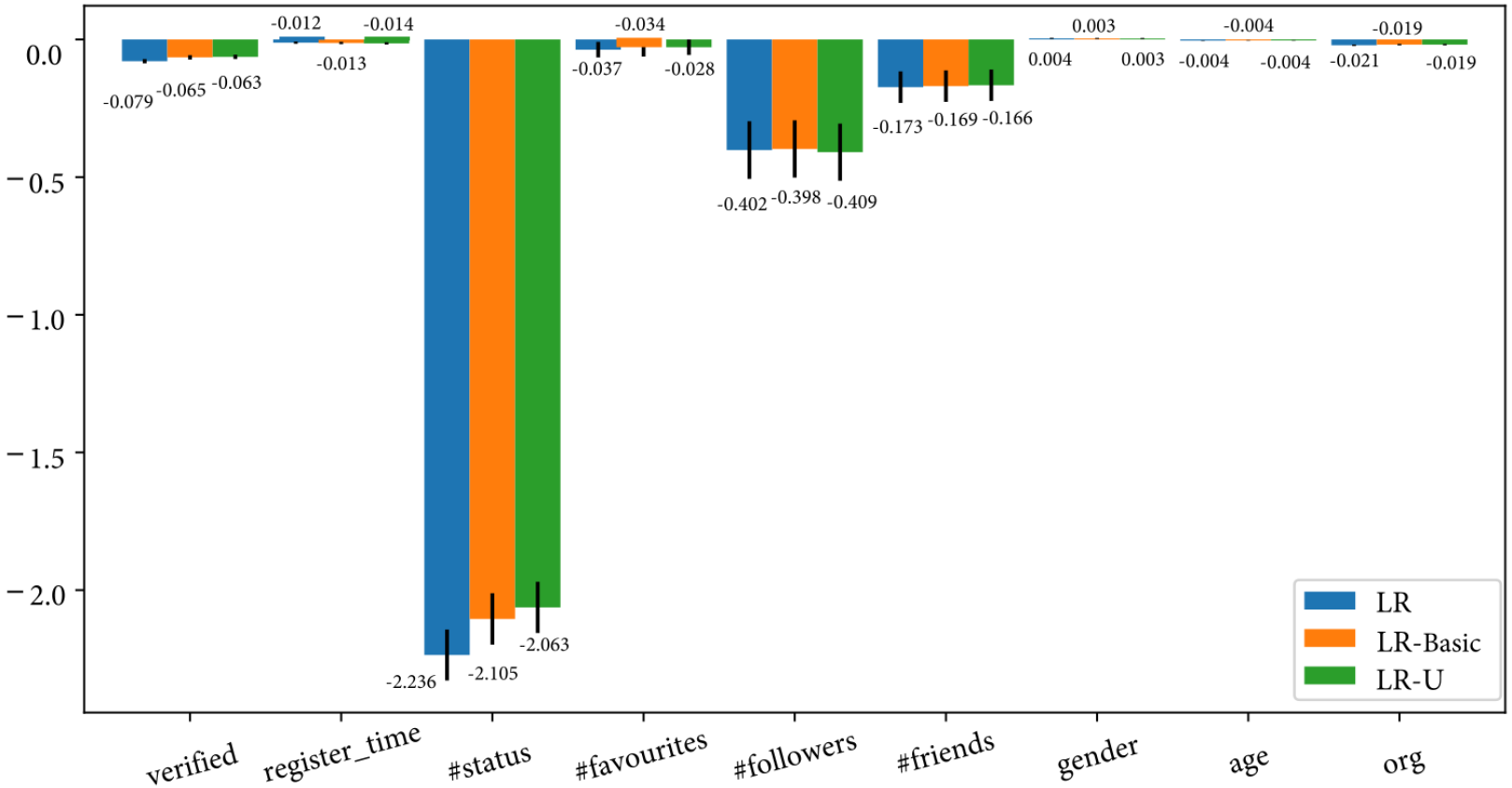


Figure 4: *PolitiFact:* Effects comparisons w.r.t. each potential causal user attribute. All the results are statistically significant.

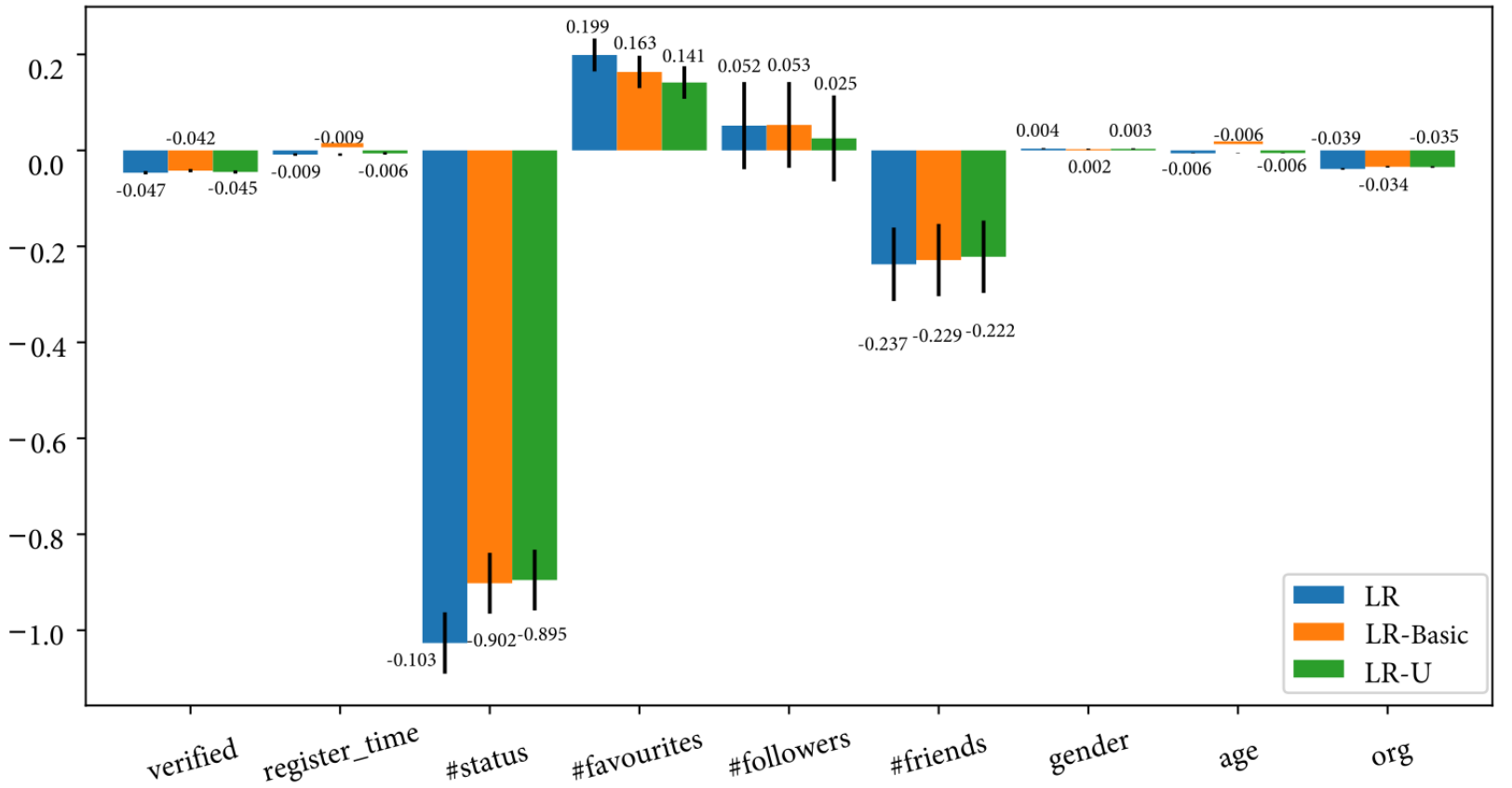


Figure 5: *GossipCop:* Effects comparisons w.r.t. each potential causal user attribute. All the results except for that of *#followers* are statistically significant.

Conclusion

1. IPS-weighted models 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
2. The identified causal attributes show that ***verified, statuses count, friends count, and org*** relate significantly with user susceptibility to share fake news.

References

- https://arxiv.org/pdf/2010.10580.pdf
- https://github.com/GitHubLuCheng/Causal-Understanding-of-Fake-News-Dissemination
- https://towardsdatascience.com/causality-an-introduction-f8a3f6ac4c4a
- https://towardsdatascience.com/causal-inference-962ae97cefda
- https://towardsdatascience.com/causal-discovery-6858f9af6dcb
- https://towardsdatascience.com/propensity-scores-and-inverse-probability-weighting-in-causal-inference-97aa53f3b6ce