PAPER TITLE

Do Graph Neural Networks Build Fair User Models? *Assessing Disparate Impact and Mistreatment in Behavioural User Profiling*

FIRST AUTHOR

Erasmo Purificato
erasmo.purificato@ovgu.de

Different **user profiling paradigms** in Graph Neural Network models **affect fairness results**

CO-AUTHOR
Ludovico Boratto

CO-AUTHOR
Ernesto W. De Luca

Motivation

- **ML models** are trained on historical data and prone to **learn biases**;
- **GNNs** can even **amplify** that **discrimination** due to the topology of graph structures;
- Only a **few publications** to evaluate **fairness** on **GNNs** and none of them consider user profiling tasks.

Contributions

- Performed two **user profiling tasks** on **real-world datasets** by using the two **most performing GNNs** in this context;
- Assessed **disparate impact** and **disparate mistreatment** with four fairness metrics;
- Correlated the different **user profiling paradigms** with the **fairness metrics** scores.

GNN models analysed

- **CatGCN** [1]: Graph Convolutional Network (GCN) model for categorical features;
- **RHGN** [2]: Relation-aware Heterogeneous Graph Network.

Fairness metrics adopted

- **Statistical parity**

$$\Delta_{SP} = |P(\hat{y}=1|s=0) - P(\hat{y}=1|s=1)|$$

- **Equal opportunity**

$$\Delta_{EO} = |P(\hat{y}=1|y=1,s=0) - P(\hat{y}=1|y=1,s=1)|$$

- **Overall accuracy equality**

$$\Delta_{OAE} = |P(\hat{y}=0|y=0,s=0) + P(\hat{y}=1|y=1,s=0) - P(\hat{y}=0|y=0,s=1) - P(\hat{y}=1|y=1,s=1)|$$

- **Treatment equality**

$$\Delta_{TE} = \left|\frac{P(\hat{y}=1|y=0,s=0)}{P(\hat{y}=0|y=1,s=0)} - \frac{P(\hat{y}=1|y=0,s=1)}{P(\hat{y}=0|y=1,s=1)}\right|$$

Experimental results

Dataset	Label	Sensitive attribute	Model	Fairness scores			
				Δ_{SP}	Δ_{EO}	Δ_{OAE}	Δ_{TE}
Alibaba	gender	bin-age	CatGCN	0.046	0.147	0.175	0.068
			RHGN	**0.018**	**0.133**	**0.148**	**0.017**
JD	gender	bin-age	CatGCN	0.033	0.050	0.062	0.150
			RHGN	**0.009**	**0.041**	**0.054**	**0.019**

Dataset	Variations in fairness scores			
	Δ_{SP}	Δ_{EO}	Δ_{OAE}	Δ_{TE}
Alibaba	0.028	0.014	0.027	**0.051**
JD	0.024	0.009	0.008	**0.131**

Observation 1. *The ability of RHGN to represent users through multiple interaction modelling gains better values in terms of fairness than a model only relying on binary associations between users and items, as CatGCN.*

Observation 2. *Even though RHGN demonstrates to be a fairer model than CatGCN, a debiasing process is equally needed for both GNNs.*

Observation 3. *In scenarios where the correctness of a decision on the target label w.r.t. the sensitive attributes is not well defined, or where there is a high cost for misclassified instances, a complete fairness assessment should always take into account disparate mistreatment evaluation.*

[1] W. Chen et al. *CatGCN: Graph Convolutional Networks with Categorical Node Features*. IEEE Trans. on Knowledge and Data Engineering (2021).
[2] Q. Yan et al. *Relation-aware Heterogeneous Graph for User Profiling*. Proc. of the 30th Int. Conf. on Information & Knowledge Management (CIKM'21).

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG — FACULTY OF COMPUTER SCIENCE
Digital Transformation and Digital Humanities — Otto von Guericke University Magdeburg
LEIBNIZ INSTITUTE FOR EDUCATIONAL MEDIA | Georg Eckert Institute
Università degli Studi di Cagliari