Compositional Fairness Constraints for Graph Embeddings*

*Joint work with my PhD supervisor Will Hamilton, to appear in ICML 2019 (pdf)

But what about fairness and privacy?

- Graph embeddings designed to capture everything that might be useful for the objective.

- Even if we don't provide the model information about sensitive attributes (e.g., gender or age), the model will use this information.

- What if a user doesn't want this information used?

Fairness in graph embeddings

- Basic idea: How can we learn node embeddings that are invariant to particular sensitive attributes?

- Challenges:
  - Graph data is not i.i.d.
  - There is not just one classification task that we are trying to enforce fairness on.
  - There are often many *possible* sensitive attributes.

Preliminaries and set-up

- Learning an encoder function to map nodes to embeddings:

$$\boxed{\mathbf{z}_v = \text{ENC}(v)}$$

- Using these embeddings to “score” the likelihood of a relationship between nodes:

$$\boxed{s(e) = s(\langle \mathbf{z}_u, r, \mathbf{z}_v \rangle)} \qquad \boxed{s(e) > s(e'), \forall e \in \mathcal{E}, e' \in \bar{\mathcal{E}}.}$$

Score of a (possible) edge is a function of the two node embeddings and the relation type.

Goal: Train the embeddings (with a subset of the true edges) so that the score for all real edges is larger than all non-edges.

Preliminaries and set-up

- Generic loss function:

$$\sum_{e \in \mathcal{E}_{\text{train}}} L_{\text{edge}}(s(e), s(e_1^-), ..., s(e_m^-))$$

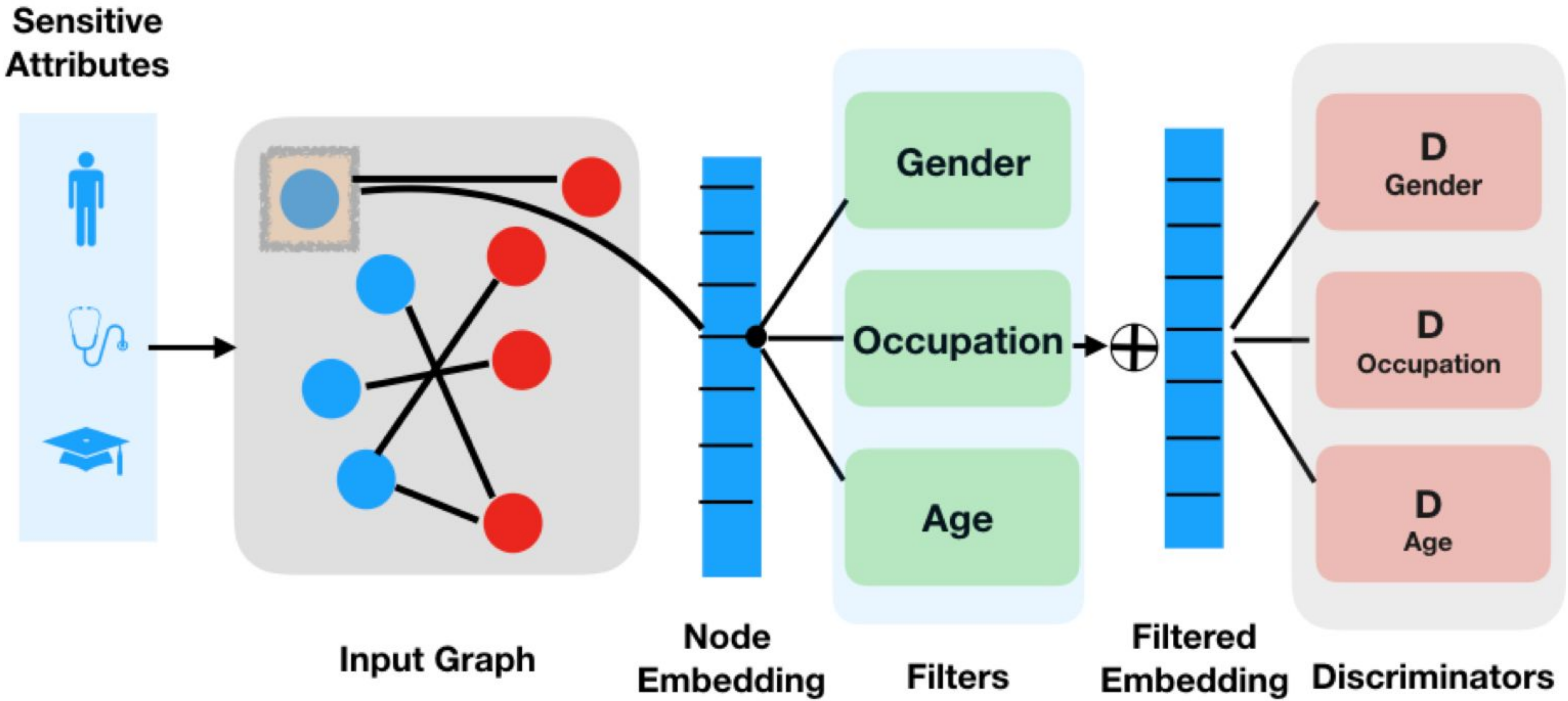
Sum over (batch of) training edges.

Task-specific loss function

Score assigned to positive/real edge.

Scores assigned to random negative sample edges.

Our work: Fairness in graph embeddings



MovieLens: Fairness results

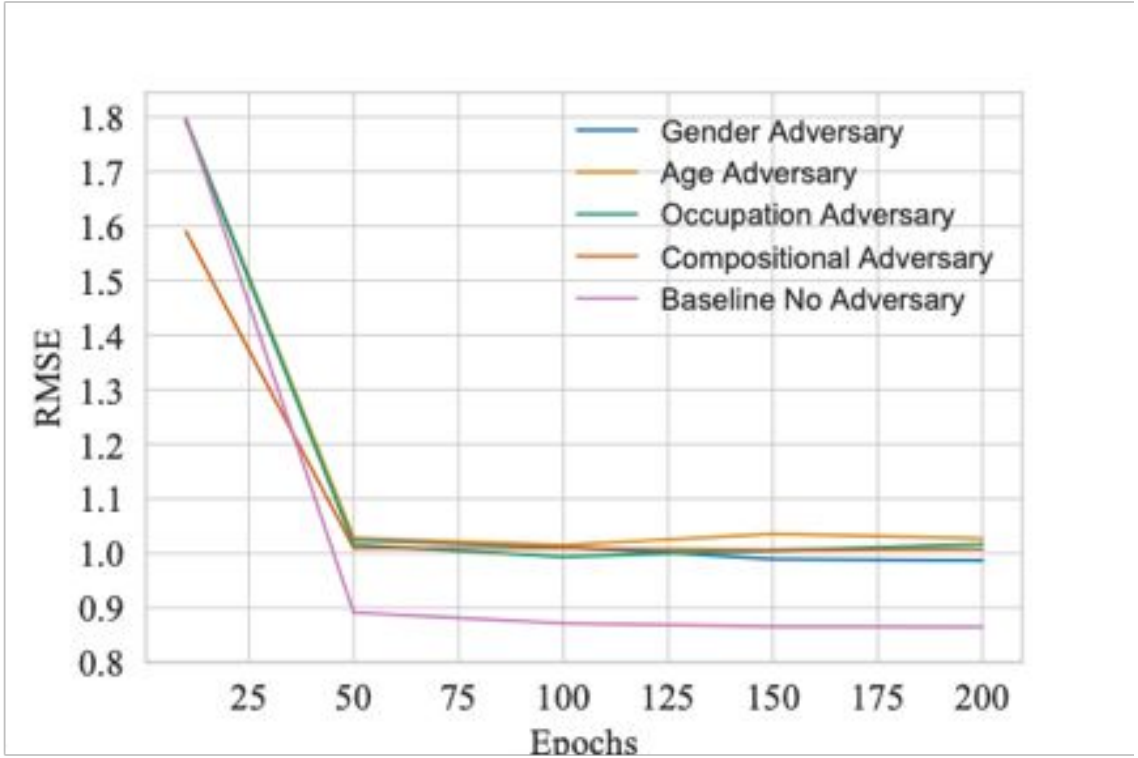
- How strongly can we enforce fairness?
- Compare three approaches to enforcing fairness:
  - No adversary (i.e., just train on the recommendation task)
  - Independent adversarial model for each attribute
  - Full compositional model

MovieLens1M	Baseline No Adversary	Gender Adversary	Age Adversary	Occupation Adversary	Comp. Adversary	Majority Classifier	Random Classifier
Gender	0.712	0.532	0.541	0.551	0.511	0.5	0.5
Age	0.412	0.341	0.333	0.321	0.313	0.367	0.141
Occupation	0.146	0.141	0.108	0.131	0.121	0.126	0.05

MovieLens: Impact on recommendations

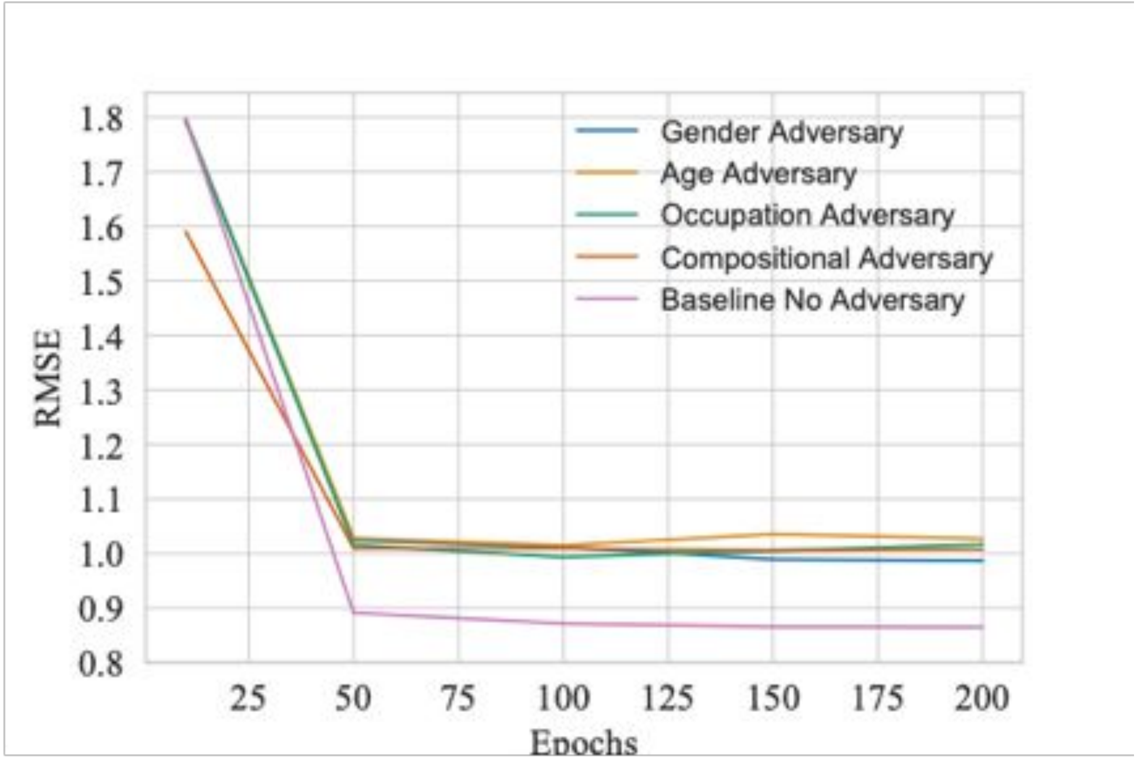
- Evaluate recommendation performance (RMSE) with and without enforcing fairness.

- There is a drop in accuracy, but not catastrophic.



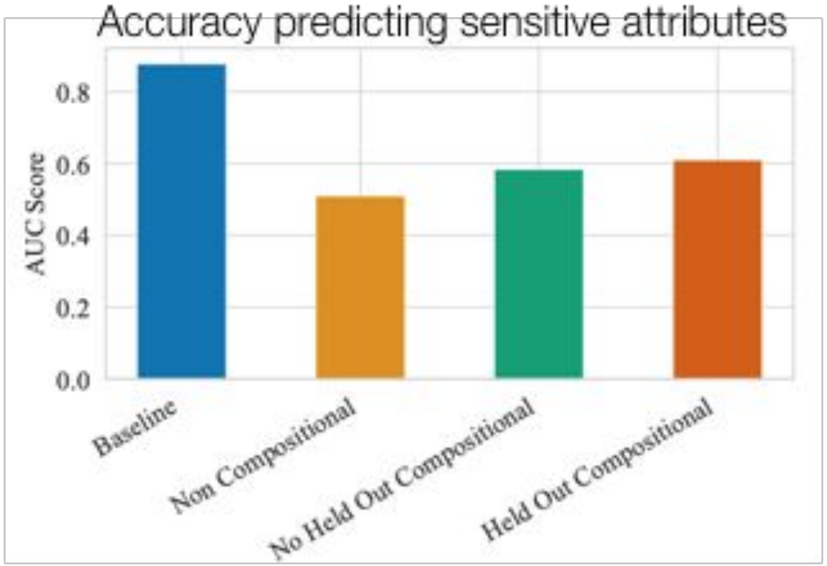
MovieLens: Impact on recommendations

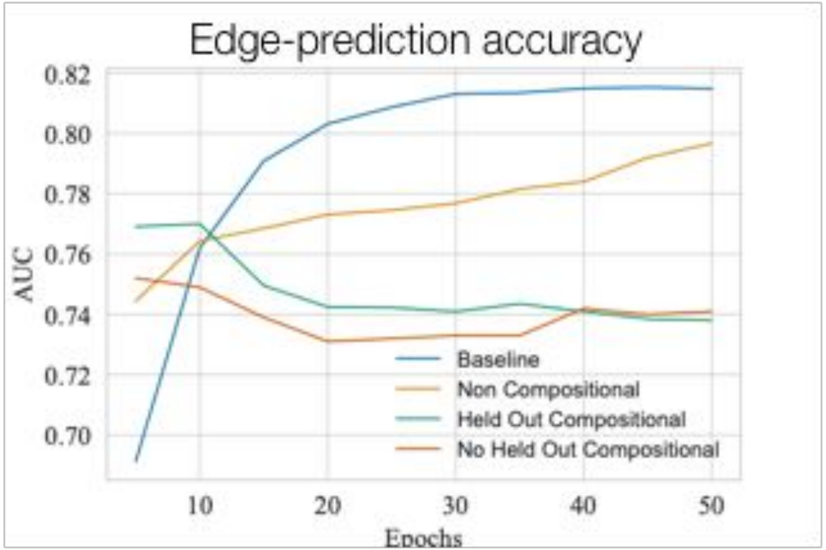
- Evaluate recommendation performance (RMSE) with and without enforcing fairness.
- There is a drop in accuracy, but not catastrophic.



Reddit results: Fairness

- Same set-up as MovieLens, but here we have 10 sensitive attributes.
- Again, able to strongly enforce fairness, but at a non-trivial cost.





Conclusions and outlook

- Fairness in network representation learning is an understudied issue.

- We can enforce fairness in a flexible way, but at a cost.

- There is no perfect notion of fairness.

Poster: Pacific Ballroom #178