Supplementary Material

A. Implementation Details

We implement each discriminator and adversarial filter as multi-layer perceptrons (MLPs) with a leaky ReLU nonlinearity between layers, and we use the Adam optimizer with default parameters. Unless otherwise specified we use $\lambda = 1000$ for all experiments and datasets. For fair comparison our discriminator during training time and subsequent sensitive attribute classifier share the same architecture and capacity. Finally, for every step performed by the main encoding model the Discriminator is updated 5 times. We found that this was necessary to provide a sufficient supervisory signal to the main encoding model.

B. FB15k-237 Details

To generate negative triplets we randomly sample either a head or tail entity during training, with a ratio of 20 negatives for each positive triplet. The TransD model is trained for 100 epochs with an embedding dimension of 20, selected using cross-validation, while the sensitive attribute classifers are trained for 50 epochs. The discriminators, sensitive attribute classifier and adversarial filters are modelled as MLP's with 4,4 and 2 layers respectively. Lastly, we use the training, validation and testing splits provided in the datasets.

C. MovieLens1M

As with FB15k-237 we use model the discriminators and sensitive attribute classifiers are modelled as MLP's but 9 layers with dropout with $p = 0.3$ between layers while the adversarial filter remains unchanged from FB15k-237. We found that regularization was crucial to the performance of main model and we use BatchNorm after the embedding lookup in the main model which has an embedding dimensionality of 30. As only user nodes contain sensitive attributes our discriminators do not compute losses using movie nodes. Finally, to train our sensitive attribute classifier we construct a 90% split of all users while the remaining user nodes are used for test. The same ratio of train/test is used for the actual dataset which constains users,movies and corresponding ratings for said movies. Finally, we train the main model and sensitive attribute classifiers for 200 epochs.

Table 1. Average AUC values across top-k sensitive attributes for Reddit. The results are reported on a Held Out test of different combinations of attributes.

REDDIT	HELD OUT AUC
20 SENSITIVE ATTRIBUTES	0.569
30 SENSITIVE ATTRIBUTES	0.569
40 SENSITIVE ATTRIBUTES	0.556
50 SENSITIVE ATTRIBUTES	0.519

D. Reddit

Like FB15k-237 we generate negative triplets by either sampling head or tail entities which are either users or subreddits but unnlike FB15k-237 we keep the ratio of negatives and positives the same. We also inherit the same architectures for discriminator, sensitive attribute classifier and attribute filters used in MovieLens1M. The main model however uses an embedding dimensionality of 50. Similar to MovieLens1M only user nodes contain sensitive attributes and as such the discriminator and sensitive attribute classifier does not compute losses with respect to subreddit nodes. Also, our training set comprises of a 90% split of all edges while the the remaining 10% is used as a test set. To test compositional generalizability we held out 10% of user nodes. Lastly, we train the main model for 50 epochs and the sensitive attribute classifier for 100 epochs.

E. Additional Results on Reddit

To the test degree of which invariance is affected by the number of sensitive attributes we report additional results on the Reddit dataset. Specifically, we report results for the Held out set with 20, 30, 40, and 50 sensitive attributes. Overall, these results show no statistically significant degradation in terms of invariance performance or task accuracy.