Fairwalk: Towards Fair Graph Embedding

Tahleen Rahman[*] , **Bartlomiej Surma**[*] , **Michael Backes** and **Yang Zhang**[†]
CISPA Helmholtz Center for Information Security
{tahleen.rahman, bartlomiej.surma, backes, yang.zhang}@cispa.saarland

Abstract

Graph embeddings have gained huge popularity in the recent years as a powerful tool to analyze social networks. However, no prior works have studied potential bias issues inherent within graph embedding. In this paper, we make a first attempt in this direction. In particular, we concentrate on the fairness of node2vec, a popular graph embedding method. Our analyses on two real-world datasets demonstrate the existence of bias in node2vec when used for friendship recommendation. We therefore propose a fairness-aware embedding method, namely *Fairwalk*, which extends node2vec. Experimental results demonstrate that *Fairwalk* reduces bias under multiple fairness metrics while still preserving the utility.

1 Introduction

The rapid assimilation of online social networks (OSNs) into people's daily lives has resulted in a massive wealth of user generated data. Researchers have created various tools to mine this rich and diverse data. One of the most prominent tools in this domain is graph embedding, which maps each user into a lower dimension vector that reflects the user's structural information within the network, such as her neighborhood, communities she belongs to, her popularity, etc. Embedding vectors have been used in various tasks, such as friendship recommendation and personal attribute prediction.

Most network embedding methods rely on deep neural networks and are often treated as a black box. The resulting vectors may have captured undesired sensitive information, that reinforces the bias already existing in the network. When used for more advanced tasks, e.g., friendship recommendation, they may result in unanticipated fairness issues. Indeed, different groups based on ethnicity, gender, levels of urbanization, or wealth are usually unequally represented in the social network graph. Analyzing bias issues inherent in such embedding based algorithms is therefore especially important but has not received attention from academia so far.

In this paper, we take the first step towards quantifying and addressing fairness issues of graph embedding methods. In particular, we concentrate on one of the most prominent methods in this field, namely node2vec [Grover and Leskovec, 2016], and investigate its fairness with respect to friendship recommendation. As pointed in [Stoica *et al.*, 2018], a recommendation system biased towards a majority in an OSN can prevent minorities from rising in the network (becoming influencers with high reach). This is further confirmed by studies on strongly homophilic networks and their negative implications on minorities [Karimi *et al.*, 2018].

Anti-discrimination laws in various countries prohibit unfair treatment based on certain traits such as race, religion, gender (sensitive attributes). The Title VII of the Civil Rights Act of 1964 is one of the most noteworthy examples. However, these laws typically are too abstract for computation, hence there exist a variety of interpretation and the resulting proposals for mathematical formulations. In this paper, we focus on disparate impact also known as group fairness [Barocas and Selbst, 2016]. Disparate impact occurs when the decision of an algorithm benefits or hurts (a) certain sensitive feature group(s) more frequently than other group(s).

1.1 Contribution

First, we conduct the first-of-its-kind study of algorithmic fairness in the setting of graph embedding methods. We specifically analyze group fairness (disparate impact) of the state-of-the-art graph embedding: node2vec.

Second, we extend *statistical parity*, a well-known measure of disparate impact to measure fairness for groups based on sensitive attributes of pairs of users. We further propose a novel notion of *Equality of Representation* to measure fairness in friendship recommendation systems.

Third, we apply node2vec for friendship recommendation in real world OSN datasets as a case study. Using the fairness metrics above, we find biases in the recommendations caused by unfair graph embeddings.

Fourth, to mitigate the aforementioned biases, we propose a novel fairness-aware graph embedding algorithm *Fairwalk*, that extends node2vec. Evaluations on our OSN demonstrate the effectiveness of *Fairwalk* on both datasets w.r.t. all fairness metrics. Compared to node2vec, *Fairwalk* increases fairness in each case by a large margin. We also compare the utility of *Fairwalk* and node2vec.

[*] Authors contributed equally to this work.

[†] Corresponding author

2 Related Work

Algorithmic bias has received a lot of attention in the recent years [Barocas and Selbst, 2016; Romei and Ruggieri, 2014]. Prior work distinguishes between group unfairness and individual unfairness. Group unfairness (disparate impact) occurs when the decision outcomes disproportionately benefit or hurt people of different groups based on their sensitive attribute value. Examples of group fairness are statistical parity [Feldman *et al.*, 2015], disparate mistreatment [Zafar *et al.*, 2017], equality of opportunity [Hardt *et al.*, 2016] and calibration [Kleinberg *et al.*, 2017]. Individual unfairness occurs when the decision of an algorithm is different for two individuals having the same non-sensitive attributes but different values for the sensitive attribute [Dwork *et al.*, 2012].

Different fairness-aware algorithms have been proposed to achieve group and individual fairness, mostly for predictive tasks. [Calders and Verwer, 2010] extend naive Bayes models by modifying the learning algorithm. [Kamiran *et al.*, 2010] modify the entropy-based splitting criterion in decision tree induction to account for sensitive attributes. [Kamishima *et al.*, 2012] apply a regularization to probabilistic discriminative models, such as logistic regression. Other works include fairness in OSN timelines [Hargreaves *et al.*, 2019] and extension of group fairness to fair rankings, [Zehlike *et al.*, 2017; Yang and Stoyanovich, 2017]. Contrary to the above, we focus on obtaining fair features that can later be used for plethora of machine learning tasks. We tackle group fairness and not individual fairness since two individuals that have identical neighborhoods would have similar embeddings, independent of the sensitive attributes of their neighbors and would be treated equally.

In the area of graph embeddings, after invention of word2vec [Mikolov *et al.*, 2013], many prominent graph data mining techniques were developed adopting the skip-gram model, namely DeepWalk [Perozzi *et al.*, 2014], node2vec [Grover and Leskovec, 2016], LINE [Tang *et al.*, 2015b] and PTE [Tang *et al.*, 2015a], and walk2friends [Backes *et al.*, 2017]. In this paper we study the fairness of node2vec - the most widely used one[*].

3 Fairness Metrics

In this section, we present the notations used throughout the paper, followed by definitions of the fairness metrics. We focus on the notion of disparate impact. Firstly, we leverage the most well-known measure of disparate impact, namely *statistical parity*. Secondly, since our work is aimed at improving the representation of under-represented groups, we propose our own notion of *Equality of Representation* which comes in two variants - the user and the network level.

3.1 Notations

We define an OSN as an undirected, unweighted graph $\mathbb{G} = (U, E)$, where the set of vertices U represents the set of users and the set of edges $E \subseteq \{(u,v)\colon u \in U, v \in U, u \neq v\}$ represents the set of friendships. Note that since $\mathbb{G}$ is undirected $(u,v) \in E$ implies $(v,u) \in E$. We define neighbors of a node u as $\omega(u) = \{v\colon (u,v) \in E\}$.

We denote a sensitive attribute by $\mathcal{S}$, and gender as $\mathcal{S} = g$ and race as $\mathcal{S} = r$. We represent the set of all possible values of $\mathcal{S}$ by $\mathcal{Z}^{\mathcal{S}}$ and denote a specific value by $z^{\mathcal{S}} \in \mathcal{Z}^{\mathcal{S}}$. The function $\zeta^{\mathcal{S}} : U \to \mathcal{Z}^{\mathcal{S}}$ maps users to their attribute values. For example, for an Asian male u, $\zeta^{g}(u) =$ "`male`" and $\zeta^{r}(u) =$ "`asian`". We denote neighbors of u with an attribute value z as $\omega_z(u) = \{v\colon v \in \omega(u) \wedge \zeta(v) = z\}$. Note that $\bigcup_{z \in \mathcal{Z}} \omega_z(u) = \omega(u)$. We define a set of users recommended to user u as $\rho\colon U \to 2^U$ and set of users with a specific attribute value z recommended to u as $\rho_z(u) = \{v\colon v \in \rho(u) \wedge \zeta(v) = z\}$.

We partition user pairs $(u,v) \in U \times U$ into groups $G^{\mathcal{S}}_{ij}$ based on the attribute values of both u and v, specifically, for $i, j \in \mathcal{Z}^{\mathcal{S}}$, $G^{\mathcal{S}}_{ij} = \{(u,v) : \zeta(u) = i \wedge \zeta(v) = j \wedge u, v \in U\}$. We collectively refer to the set of all groups based on a sensitive attribute $\mathcal{S}$ as $\mathcal{G}^{\mathcal{S}}$.

3.2 Statistical Parity

Statistical Parity also known as *Demographic Parity* or *Independence* is the statistical equivalent of the legal doctrine of disparate impact [Feldman **et al.**, 2015]. *Statistical Parity* (typically defined in terms of two groups) requires the acceptance rates of the candidates from both groups to be equal. This allows us to measure recommendation fairness independent of the ground truth (existing friendships) which in our case is itself biased. Given a partitioning of user pairs based on an attribute $\mathcal{S}$, into two groups $G^{\mathcal{S}}_{ab}$ and $G^{\mathcal{S}}_{cd}$, let $P(G^{\mathcal{S}}_{ij})$ denote the acceptance rate for group $G^{\mathcal{S}}_{ij}$,

$$P(G^{\mathcal{S}}_{ij}) = \left|\{(u,v) : v \in \rho(u) \wedge (u,v) \in G^{\mathcal{S}}_{ij}\}\right| / \left|G^{\mathcal{S}}_{ij}\right|$$

The bias, or statistical parity, w.r.t. $G^{\mathcal{S}}_{ab}$ and $G^{\mathcal{S}}_{cd}$ is then the difference between $P(G^{\mathcal{S}}_{ab})$ and $P(G^{\mathcal{S}}_{cd})$.

We extend the above definition to account for multiple groups that we have for friendship recommendation. Since our groups consider the attribute values of both users in a pair, we have at least 4 groups when we consider a binary attribute like gender and in general $\left|\mathcal{Z}\right|^2$ groups when we consider any n-ary attribute like race. To capture the differences between multiple groups, we calculate the variance between the acceptance (recommendation) rates of each group in $\mathcal{G}^{\mathcal{S}}$.

$$\text{bias}^{\texttt{SI}}(\mathcal{G}^{\mathcal{S}}) = \text{Var}(\{P(G^{\mathcal{S}}_{ij})\} : G^{\mathcal{S}}_{ij} \in \mathcal{G}^{\mathcal{S}}) \qquad (1)$$

3.3 Equality of Representation

Modern online recommendation systems suffer from problems of echo chambers or the information bubble effect [Flaxman *et al.*, 2016; Quattrociocchi *et al.*, 2016]. Usually users get recommendations based on their interests, which isolates them from other contradicting interests or viewpoints and reinforces the existing viewpoints. This also manifests in friendship recommendations in OSNs, where users who join a network into some well-established community, rarely see anything or anybody outside of it, although it might be interesting for them. As a direct consequence minority communities get isolated and are not seen by others. Our work aims to improve the representation of such under-represented groups in the OSN graph embeddings. To promote recommendations

[*] according to citation count from GoogleScholar

where all groups are equally represented, we define bias by *Equality of Representation* in two variants: bias^{ERg} at the network level and bias^{ERu} at the user level.

Network level. At the network level, we measure bias between different groups $G^{\mathcal{S}}_{ij}$, among all recommendations given in the network. Denoting the number of recommendations from a group $G^{\mathcal{S}}_{ij}$ as $N(G^{\mathcal{S}}_{ij}) = |\{(u,v) : v \in \rho(u) \wedge (u,v) \in G^{\mathcal{S}}_{ij}\}|$,

$$\text{bias}^{\text{ERg}}(\mathcal{G}^{\mathcal{S}}) = \mathrm{Var}(\{N(G^{\mathcal{S}}_{ij})\} : G^{\mathcal{S}}_{ij} \in \mathcal{G}^{\mathcal{S}}) \tag{2}$$

User level. At the user level, among the recommendations $\rho(u)$ given to each user u, we measure the fraction of users having attribute value z and denote it as $z\text{-share}(u) = \frac{|\rho_z(u)|}{|\rho(u)|}$. Specifically, for a given attribute value z, bias is measured as the difference between a *fair fraction* (where each attribute value has an equal share) and the average z-share over all users.

$$\text{bias}^{\text{ERu}}(z) = \frac{1}{|\mathcal{Z}^{\mathcal{S}}|} - \frac{\sum_{u \in U} z\text{-share}(u)}{|U|} \tag{3}$$

This definition allows for measuring the bias for each sensitive attribute value independently.

4 Friendship Recommendation with node2vec

In this section, we first review the graph embedding algorithm node2vec, then describe a node2vec based recommendation system and finally analyze its fairness.

4.1 Graph Embedding - node2vec

A graph embedding is a mapping from nodes in a graph to a vector space $f : U \rightarrow \mathbb{R}^d$ where d is a hyperparameter capturing the number of dimensions of the vector space. Resulting vectors can be used for multiple machine learning tasks, e.g., link prediction for friend recommendation. Node2vec [Grover and Leskovec, 2016] first uses a random walk over a graph to generate walk traces and then extracts features based on learned traces.

Random walk. Given a graph $\mathbb{G}$, for each node $u \in U$, node2vec performs a single random walk `walk_num` number of times. The result is a list of traces, where each trace is a list of nodes' ids resulting from single random walks. Single random walk is a standard random walk over graphs without weights, i.e., at each step the next node is chosen uniformly at random among all neighbors of the current node. Both `walk_num` and `walk_len` are hyperparameters of node2vec.

Feature learning. Next, node2vec uses the generated traces to train a neural network and learn embedding vectors. Let us define a *network neighborhood* $N_s(u)$ as the set of nodes preceding and succeeding node u in the generated walk traces. The skip-gram architecture is adapted to the network traces with a goal to maximize the log-probability of observing a *network neighborhood* $N_s(u)$ based on the feature vector of node u. The corresponding objective function is defined as:

$$\arg\max_{f} \prod_{u \in U} \prod_{u' \in N_s(u)} P(u'|f(u)) \tag{4}$$

Here, the conditional probability $P(u'|f(u))$ is modeled as a softmax function.

4.2 Recommendation System

We take a supervised learning approach towards the recommendation task. For each user, we use the graph embedding as described above in 4.1 to learn embeddings. We further define a feature vector $\vec{x}_{(u,v)}$ for a user pair (u,v) as the Hadamard distance[†] between embeddings of u and v. Given two vectors $f(u)$ and $f(v)$, the Hadamard operator $\boxdot$ is defined as: $[f(u) \boxdot f(v)]_i = [f(u)]_i[f(v)]_i$. We train a random forest classifier to learn associations between the feature vectors of user pairs $\vec{x}_{(u,v)}$ and presence or absence of friendships between them.

The trained model is then used to predict class probabilities for unseen candidate pairs using their corresponding feature vectors $\vec{x}_{(u,v)}$. We use the positive class probability as the recommendation score. A friend v is recommended to a user u, if the recommendation score for the pair $(u,v) \notin E$ is within the top $k\%$ of all the scores received by all candidate pairs. We denote the resulting friendship recommendations given to a user u by $\rho(u)$.

4.3 Fairness of node2vec

We now evaluate the fairness of node2vec-based friendship recommendation using *Equality of Representation* at user level, bias^{ERu} (Eq. (3)). In a fair recommendation system, we would expect the bias to be lower than the initial bias. To calculate the initial bias in the network, we modify Eq. (3) by using the neighborhood function ω instead of recommendation ρ. We do a rigorous bias evaluation using all metrics in Section 6.

We distinguish between two genders $\mathcal{Z}^g = \{z^g_0, z^g_1\}$ and three races: $\mathcal{Z}^r = \{z^r_0, z^r_1, z^r_2\}$ and the protected (minority) groups are: z^g_1, z^r_0, z^r_2 (we describe our dataset in details in Section 6.1). Figure 1 shows the distribution of $z\text{-share}$ values for all users. The distribution of $z\text{-share}$ indeed follows the original distribution in the network, and so does the bias. This demonstrates that node2vec based recommendations mirrors the gap between minorities and majorities.

5 Fairwalk

Since graph embeddings aim to find the best structural representation of nodes, the algorithm also unintentionally learns some information about people's sensitive attributes and relations between them. Recommendations based on this further reinforces differences between minorities and majorities. As a countermeasure to this problem, we propose *Fairwalk*, a modified version of random walk, which results in a more diverse *network neighborhood* representation thereby producing less biased graph embedding.

5.1 Random Walk

We are modifying the random walk procedure from original node2vec. Instead of randomly selecting a node to jump to

[†] While other binary operators (e.g., average, L1, L2 distance) could be used, we chose Hadamard distance, as it performed the best in original node2vec paper [Grover and Leskovec, 2016]

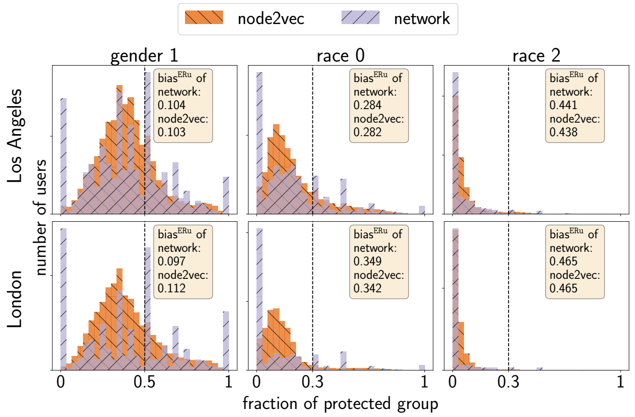

Figure 1: z-share distributions of node2vec and original network. The vertical line shows the *fair fraction* (0.5 and 0.3)

from amongst all neighbors, we now partition neighbors into groups based on their sensitive attribute values and give each group the same probability of being chosen regardless of their sizes. Then a random node from the chosen group is selected for the jump. The modified random walk procedure is shown in Algorithm 1. By $\xleftarrow{R} \mathbb{A}$ we denote drawing an element of set $\mathbb{A}$ uniformly at random.

Algorithm 1 Fair random walk trace generation

```
procedure RAND_WALK($U, \omega$, walk_num, walk_len)
    traces $\leftarrow$ empty_list
    for all $u \in U$ do
        for $i \leftarrow 0$, walk_num do
            trace $\leftarrow$ empty_list
            $u_1 \leftarrow u$
            for $j \leftarrow 0$, walk_len do
                trace.append($u_1$)
                $\mathcal{Z}_u \leftarrow \{z : z \in \mathcal{Z} \wedge |\omega_z(u_1)| > 0\}$
                $z_1 \xleftarrow{R} \mathcal{Z}_u$
                $v \xleftarrow{R} \omega_{z_1}(u_1)$
                $u_1 \leftarrow v$
            end for
            traces.append(trace)
        end for
    end for
    return traces
end procedure
```

5.2 Resulting Traces

While detailed evaluation of the resulting modified embedding is in Section 6, we can already discuss differences in the walk traces themselves. From Figure 2 we can see that minorities appear more often in the fair random walk. Not only does it result in higher *network neighborhood* diversity with respect to sensitive attribute, but also minorities appear more frequently in the traces. This provides more *network neighborhood* data for minorities which enables the neural network give more importance to obtain their best vector representation while optimizing the overall objective function.

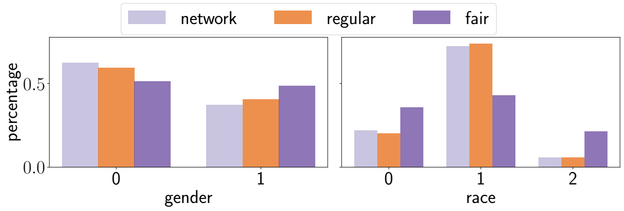

Figure 2: Ratio of each gender and race in the original network and regular and fair random walk traces in Los Angeles dataset

6 Evaluation

In this section, we first describe our dataset, followed by the set-up of the machine learning model for our experiments. We then present the evaluation results for our *Fairwalk* using the fairness metrics defined in Section 3 and finally show the recommendation utility in terms of precision and recall.

6.1 Dataset

We use Instagram data collected from two of the biggest cities in different English speaking countries, namely London and Los Angeles (LA). The data was collected in 2016 using the Instagram API. An edge exists between two users if they mutually follow each other (e.g., [Cho *et al.*, 2011]). We concentrate on two sensitive attributes, gender and race, which we extract by querying Face++ with users' profile pictures. Tools such as Face++ have limitations [Buolamwini and Gebru, 2018]. To this end, we choose users for whom the attribute with highest confidence is at least 20 percentage points higher then the second best. We distinguish between female or male gender, and denote them as 0 or 1, i.e., $\mathcal{Z}^g = \{0, 1\}$, and between African, Caucasian or Asian race and denote them as 0, 1 or 2, i.e., $\mathcal{Z}^r = \{0, 1, 2\}$. Table 2 shows the sizes of datasets and the proportions of different genders and races. We thus have 4 pairwise groups for gender namely $G_{z_0,z_0}, G_{z_0,z_1}, G_{z_1,z_0}, G_{z_1,z_1}$ and 9 pairwise groups for race namely $G_{z_i,z_j}, i, j \in \{0, 1, 2\}$. Table 1 shows the proportions of friendships in different groups.

6.2 Experimental Setup

We iterate our experiments 5 times. To this end we divide our dataset into 5 equal slices. We train a random forest with 100 trees using 4 out of 5 slices, leaving out a different slice each time. We use Hadamard distance between the feature embeddings of each user pair as input features to the random forest. Graph embeddings are trained 5 times, each time with the same 4 slices as used for training. We use the following hyper-parameters (following [Grover and Leskovec, 2016]), i.e., length of each walk: walk_len $= 80$, number of walks starting from each node in the graph: walk_num $= 20$, number of dimensions of the resulting vector space: $d = 128$. For the recommendation candidates, of each node u we rank all non-friend nodes by the cosine similarity of their embeddings with the embedding of u and select the top 100. This approach gives us a candidate set likely to be recommended.

	Gender groups				Race groups								
$i - j$ for G_{ij}	0-0	0-1	1-0	1-1	0-0	0-1	1-0	1-1	0-2	1-2	2-0	2-1	2-2
LA	37.78	21.00	21.99	19.21	5.97	12.55	12.65	57.92	1.17	3.87	1.16	3.74	0.96
London	38.96	18.19	20.74	22.10	3.55	9.52	9.83	70.31	0.47	2.57	0.56	2.76	0.41

Table 1: Percentage of existing friendships in each group in our original dataset

	LA	London
No. users	82,607	53,902
No. social links	482,305	165,184
gender 0	62.6%	62.3%
gender 1	37.4%	37.7%
race 0	21.9%	15.9%
race 1	72.2%	80.7%
race 2	5.9%	3.4%

Table 2: Statistics of both datasets.

		LA		London	
		gender	race	gender	race
ERg	regular	$1.3e^{10}$	$2.5e^{7}$	$6.5e^{9}$	$2.4e^{7}$
	fair	$0.8e^{10}$	$1.9e^{7}$	$4.8e^{9}$	$1.9e^{7}$
SI	regular	$4.7e^{-9}$	$1.4e^{-12}$	$1.1e^{-8}$	$7.1e^{-11}$
	fair	$1.7e^{-9}$	$0.4e^{-12}$	$0.2e^{-8}$	$2.8e^{-11}$

Table 3: bias^{SI} and bias^{ERg} for both cities (lower, the better)

For scalability reasons (for bigger datasets), non-friend pairs can be sampled randomly. For the quantity of top recommendations, we experiment with a variety of values for k and do not observe any significant differences. Therefore we show results below only for $k = 20$ i.e., when the top 20% of all candidates are selected for recommendation. We denote node2vec based recommendations by "regular" and *Fairwalk* based recommendation by "fair".

6.3 Statistical Imparity

Figure 3 shows the acceptance rates $P(G_{ij}^{\mathcal{S}})$(fraction of recommended users pairs out of all possible pairs in each group) for both the regular and the *Fairwalk*. We observe that, *Fairwalk* increases the probability of the under represented groups being recommended to a very large extent. Additionally, it is noteworthy to see that the probabilities of same gender and same race friendship recommendations are always reduced by *Fairwalk* and the probabilities of diverse friendships are always increased.

Table 3 shows bias^{SI} compared to node2vec for both cities. *Fairwalk* reduces $\text{bias}^{\text{SI}}(\mathcal{G}^g)$ by 61% and $\text{bias}^{\text{SI}}(\mathcal{G}^r)$ by 68% for LA. For London, *Fairwalk* reduces $\text{bias}^{\text{SI}}(\mathcal{G}^g)$ by 91% and $\text{bias}^{\text{SI}}(\mathcal{G}^r)$ by 61% compared to node2vec.

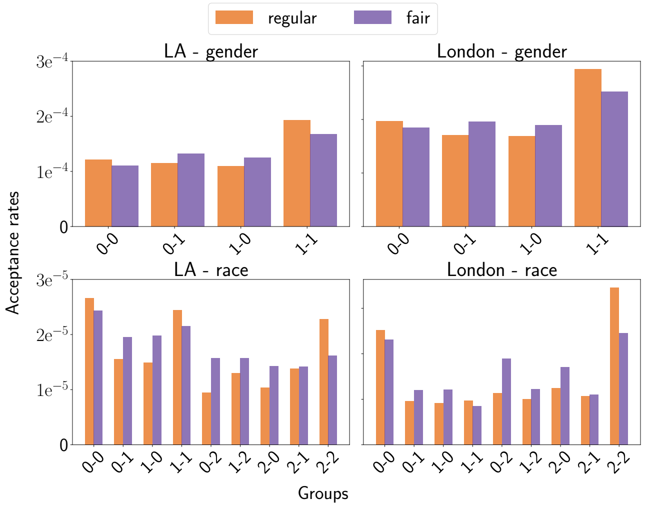

Figure 3: Fraction of recommended users pairs out of all possible pairs in each group. The x-axes marks the Type-2 groups G_{z_i,z_j} with the corresponding $i - j$

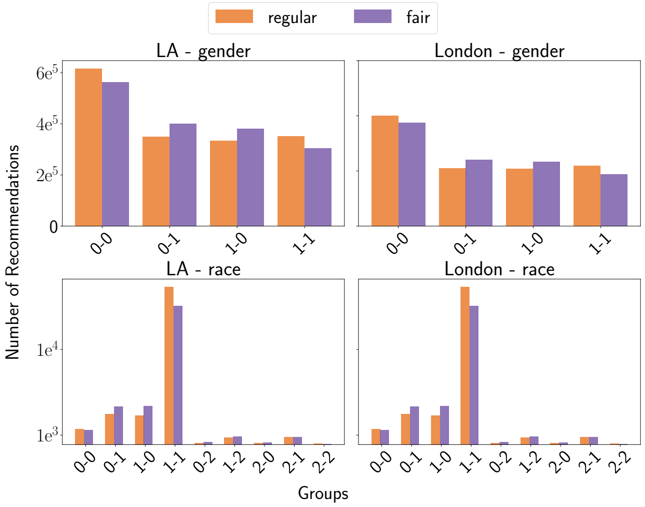

Figure 4: Number of recommended users pairs from each group. The x-axes marks groups G_{ij} with the corresponding $i - j$

6.4 Equality of Representation

Group Level. Figure 4 highlights the differences between the representations of different groups among the top recommendations. We see that as compared to node2vec, *Fairwalk* decreases the share for the over-represented groups, e.g., $\mathcal{G}_{11}^r$

		gender		race		
		0	1	0	1	2
LA	network	0.104	0.104	0.117	0.392	0.275
	node2vec	0.103	0.103	0.115	0.387	0.272
	fairwalk	0.068	0.068	0.054	0.288	0.234
London	network	0.097	0.097	0.183	0.481	0.298
	node2vec	0.112	0.112	0.176	0.474	0.298
	fairwalk	0.095	0.095	0.135	0.417	0.282

Table 4: Bias by *Equality of Representation* at user level for both genders and all three races (lower, the better).

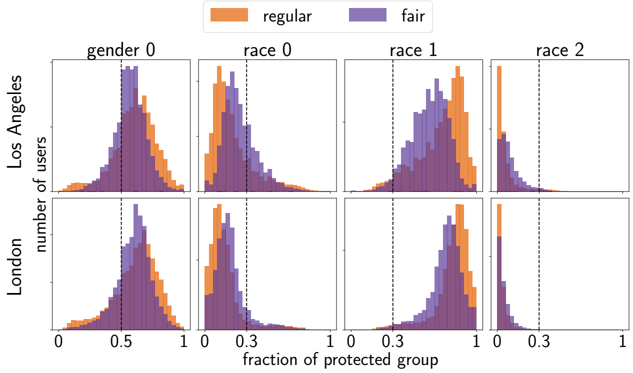

Figure 5: z-share distributions of node2vec and *Fairwalk*. The vertical line shows the *fair fraction*.

and $\mathcal{G}^g_{00}$ and increases the share for most underrepresented groups which can be clearly seen for $\mathcal{G}^r_{01}$, $\mathcal{G}^r_{10}$, $\mathcal{G}^r_{02}$, $\mathcal{G}^r_{12}$ and $\mathcal{G}^g_{10}$, $\mathcal{G}^g_{01}$. For $\mathcal{G}^g_{11}$ and $\mathcal{G}^r_{00}$, *Fairwalk* could not increase the representation anymore since they are already very strongly connected as seen in Fig 3. Table 3 shows bias^{ERg} for both cities. Compared to node2vec, *Fairwalk* reduces $\text{bias}^{\text{ERg}}(\mathcal{G}^g)$ by 36% and $\text{bias}^{\text{ERg}}(\mathcal{G}^r)$ by 23% for LA and $\text{bias}^{\text{ERg}}(\mathcal{G}^g)$ by 26% and $\text{bias}^{\text{ERg}}(\mathcal{G}^r)$ by 21% for London.

User Level. On the distribution plots on Figures 5 we can see that *Fairwalk* reduces the gap between different groups, by leaning towards the *fair fraction* (where each attribute value has an equal share). We skipped the distribution for gender 1 since it's symmetrical to gender 0. The exact values of bias^{ERu} for different groups for the original network (initial bias) for node2vec and *Fairwalk* can be found in Table 4. We see that *Fairwalk* decreased the bias in all cases, for LA the average improvement is 32% and for London 14%. The best result is for race 0 in LA, with an improvement of 53%.

6.5 Precision and Recall

We also evaluate whether our recommendations capture any friendship edges that would otherwise have been formed naturally without users following any recommendations at all. To this end, we use the 20% of friendship edges unused during training for each iteration as described in Sec. 6.2 as ground truth. We calculate precision and recall for *Fairwalk*, and compare with the regular node2vec recommendations.

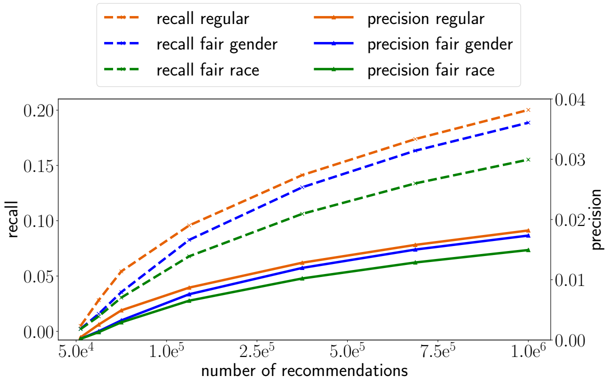

Figure 6: Precision and recall for different number of recommendations.

Figure 6 shows precision and recall for different number of recommendations for LA. Unsurprisingly, we see that precision and recall are always lower for *Fairwalk* as compared to node2vec. This indicates that we deviate more from the original biased growth of the network compared to regular recommendations thereby not amplifying the initial bias in the network. We also see that for race, *Fairwalk* deviates the most since it tries to balance more groups compared to gender, and given that race 1 alone makes up a huge proportion of our OSN dataset. Nevertheless, we capture a large number of true positives. Results for London follow the same trend.

7 Conclusion and Future Work

We study fairness issues in a state-of-the-art graph embedding method node2vec. Using the metrics of *Statistical Parity* and *Equality of Representations* we find bias in node2vec. We propose a fairness-aware *Fairwalk* and demonstrate its effectiveness in mitigating the aforementioned biases by a large scale evaluation on real world OSN datasets for friendship recommendation.

While *Fairwalk* already improves *Statistical Parity* and *Equality of Representations* to a large extent, there exist multiple other notions of fairness in the community dependent on the task at hand. Our *Fairwalk* can be tuned with parameters to fulfill any such fairness notion, for example maintaining the real-world ratio between groups. One could also iteratively perform further recommendation to achieve a higher balance or parity. Using a weighted ensemble-like approach that allow optimizing for a combination of sensitive attributes (e.g Asian females) can account for cross-attribute biases in the network. This can also be done following recent work on rich subgroup fairness [Kearns *et al.*, 2019].

Acknowledgements

The authors would like to thank Inken Hagestedt, David Pfaff as well as the anonymous reviewers for their valuable comments on the submitted manuscript.

References

Michael Backes, Mathias Humbert, Jun Pang, and Yang Zhang. walk2friends: Inferring Social Links from Mobility Profiles. In *Proc. CCS*, pages 1943–1957, 2017.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Cal. L. Rev.*, 104:671, 2016.

Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proc. FAT**, pages 77–91, 2018.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

Eunjoon Cho, Seth A. Myers, and Jure Leskovec. Friendship and Mobility: User Movement in Location-based Social Networks. In *Proc. KDD*, pages 1082–1090, 2011.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness Through Awareness. In *Proc. ITCS*, pages 214–226, 2012.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proc. KDD*, pages 259–268, 2015.

Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320, 2016.

Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks. In *Proc. KDD*, pages 855–864, 2016.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. In *Proc. NIPS*, pages 3315–3323, 2016.

Eduardo Hargreaves, Claudio Agosti, Daniel Menasché, Giovanni Neglia, Alexandre Reiffers-Masson, and Eitan Altman. Fairness in online social network timelines: Measurements, models and mechanism design. *Performance Evaluation*, 129:15–39, 2019.

Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination Aware Decision Tree Learning. In *Proc. ICDM*, pages 869–874, 2010.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Proc. ECML/PKDD*, pages 35–50, 2012.

Fariba Karimi, Mathieu Génois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. Homophily Influences Ranking of Minorities in Social Networks. *Scientific Reports*, 2018.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An Empirical Study of Rich Subgroup Fairness for Machine Learning. In *Proc. FAT**, pages 100–109, 2019.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 67. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *Proc. ICLR*, 2013.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online Learning of Social Representations. In *Proc. KDD*, pages 701–710, 2014.

Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. Echo chambers on facebook. *Available at SSRN 2795110*, 2016.

Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014.

Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity. In *Proc. WWW*, pages 923–932, 2018.

Jian Tang, Meng Qu, and Qiaozhu Mei. PTE: Predictive Text Embedding through Large-scale Heterogeneous Text Networks. In *Proc. KDD*, pages 1165–1174, 2015.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. LINE: Large-scale Information Network Embedding. In *Proc. WWW*, pages 1067–1077, 2015.

Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *SSDBM*, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proc. WWW*, pages 1171–1180, 2017.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. FA*IR: A Fair Top-k Ranking Algorithm. In *Proc. CIKM*, pages 1569–1578, 2017.