Bursting the Filter Bubble: Fairness-Aware Network Link Prediction

Farzan Masrour, Tyler Wilson, Heng Yan, Pang-Ning Tan, Abdol-Hossein Esfahanian
Michigan State University
{masrours, wils1270, yanheng, ptan, esfahanian}@msu.edu

Abstract

Link prediction is an important task in online social networking as it can be used to infer new or previously unknown relationships of a network. However, due to the homophily principle, current algorithms are susceptible to promoting links that may lead to increase segregation of the network—an effect known as filter bubble. In this study, we examine the filter bubble problem from the perspective of algorithm fairness and introduce a dyadic-level fairness criterion based on network modularity measure. We show how the criterion can be utilized as a postprocessing step to generate more heterogeneous links in order to overcome the filter bubble problem. In addition, we also present a novel framework that combines adversarial network representation learning with supervised link prediction to alleviate the filter bubble problem. Experimental results conducted on several real-world datasets showed the effectiveness of the proposed methods compared to other baseline approaches, which include conventional link prediction and fairness-aware methods for i.i.d data.

Introduction

Online social networking sites have transformed the way individuals interact and share information with each other. The wealth of social network data available also provide opportunities to mine them for a variety of business applications. For example, businesses can learn about the users' interests, sentiment, and online behavior by analyzing the social network data. The insights gained from such analysis will help businesses to increase engagement with their existing customers or connect with new customers. Despite its importance, recent studies have raised concerns about the potential biases and unintended consequences that may arise from such automated analysis.

For example, link prediction methods (Liben-Nowell and Kleinberg 2007; Al Hasan et al. 2006; Masrour et al. 2015; 2018) are commonly employed by social networking sites to encourage users to expand their social circles. "Suggested for you" on Instagram and "People you may know" on LinkedIn are two example applications of such methods. However, the rise of link prediction systems have led to an effect known as *filter bubble* (Pariser 2012), which is the reinforced segregation and narrowing diversity of information exposed to online users. If left unchecked, the filter bubble may introduce systematic biases in the network data and its subsequent analysis. For instance, Hofstra et al. (Hofstra et al. 2017) examined the ethnic and gender diversity of social relationships on Facebook and showed that those who have ample opportunities to befriend other similar users often find themselves in highly segregated networks. This is due to the *homophily* principle (McPherson, Smith-Lovin, and Cook 2001), which is the tendency of individuals to form social ties with other similar individuals in a network. As current algorithms are designed to promote links between similar users, their suggested links may exacerbate the user segregation problem.

In addition to online social networks, the filter bubble problem is also prevalent in recommender systems, which can be viewed as a link prediction task applied to a bipartite network of users and items. A recent study by Nguyen et al. (Nguyen et al. 2014) concluded that recommender systems tend to expose users to "slightly narrowing set of items over time." For example, in movie recommendation, movies from a certain genre may only be recommended to users from a specific gender. By addressing the filter bubble problem in network link prediction, the proposed method can potentially be used to alleviate the filter bubble problem in other types of recommender systems.

This paper examines the filter bubble problem for network link prediction from algorithm fairness perspective. Specifically, we consider a link prediction algorithm to be unfair if it is biased towards promoting certain types of links (e.g., those between users with similar gender or other protected attributes). As a motivating example, consider the link prediction task on professional networking sites. Certain professions, such as software engineering, tend to be dominated by men, a fact that is likely to be reflected in the link structure of the professional network. As a result, the links recommended by the site may reinforce this gender-based segregation and primarily recommend links between individuals from the same gender while recommending comparatively fewer inter-gender links. Though such a system may be able to achieve high link prediction accuracy, it may unfairly disadvantage some users. For example, a female software engineer may be treated unfairly as they are seldom recommended to other male software engineers.

Unfair practices due to the decisions generated by auto-

mated systems is a problem that has been well-documented in many application domains, including criminal justice, mortgage lending, and university admission. For example, Angwin et al. (Angwin et al. 2016) warned about the potential biases against African Americans in the software used to predict the risk score of defendants who would likely reoffend again while O'Neil (O'Neil 2017) cautioned against the manipulative marketing tactics used by for-profit colleges in online advertising that exploit vulnerable populations. These concerns have brought increasing scrutiny into the issue of fairness in machine learning algorithms. Despite their growing research, existing works are primarily focused on independent and identically distributed (*i.i.d*) data, and thus, may not be suitable for link analysis problems. For example, previous works have considered the notion of fairness either at individual (Dwork et al. 2012) or group (Hardt et al. 2016; Feldman et al. 2015) level. In contrast, this paper examines the notion of fairness at a *dyadic-level*, based on the pairwise interactions between users in a social network. Furthermore, previous approaches have considered fairness in terms of the unjust decisions against members of a specific underrepresented (protected) group. Instead, we consider fairness in terms of promoting inter-group connections in a network in order to alleviate the filter bubble problem.

There are four major contributions of this paper. First, we empirically assess the influence of protected attributes such as gender on the link structure of a network by measuring the homophily effect on several real-world network datasets. Second, we introduce *modred* as a fairness criterion for network link prediction. The metric is inspired by the well-known modularity measure (Newman and Girvan 2004) developed for network community detection. We consider the reduction in modularity measure as a way to determine whether the links predicted by an algorithm may lead to further segregation of the network. We then illustrate how the measure can be incorporated into a greedy algorithm for postprocessing the results of current link prediction algorithms. Finally, we present a novel Fairness-aware LInk Prediction (**FLIP**) framework that combines adversarial network representation learning with supervised link prediction to mitigate the filter bubble problem.

Related Work

Link prediction is a well studied problem in network analysis with various algorithms been developed over the past two decades (Al Hasan et al. 2006; Masrour et al. 2015). This includes heuristics methods that consider the pairwise similarities between nodes, where similarity is defined based on the network topology (Newman 2001; Liben-Nowell and Kleinberg 2007) or node features (Crandall et al. 2010). The main benefit of these methods is their simplicity and the fact that most of these approaches do not required training. Another class of link prediction methods employ machine learning methods, such as those based on probabilistic graphical models (Clauset, Moore, and Newman 2008), matrix factorization (Scripps et al. 2008), and supervised classification (Al Hasan et al. 2006; Wang et al.). Despite their higher accuracy, these methods often suffer from the class imbalance problem as the number of links in a network is significantly fewer than the number of non-links. Recent years have also witnessed the emergence of deep neural network methods for the link prediction task (Li et al. 2014; ; Tian et al. 2014). These methods have been shown to achieve state of the art performance.

Social networks are increasingly personalizing their content using automated machine learning techniques, which is a concern as the decisions may lead to adverse effects on the users. This is due to the so-called "filter bubble" or "echo chamber" effect (Hofstra et al. 2017; Pariser 2012) in which individuals are increasingly isolated to consuming only information that conform to their own belief system. In online social networks, the effect of filter bubble is exemplified by the recommendation decisions generated using link prediction algorithms. As link prediction algorithms are commonly used to encourage users to expand their networks, this may lead to adverse consequences such as segregation of users (Hofstra et al. 2017; Nguyen et al. 2014).

Quantifying fairness has been a subject of intense debate among AI and ML researchers in recent years (Berk et al. 2018; Dwork et al. 2012; Hardt et al. 2016; Kusner et al. 2017). Previous works are primarily focused on non-relational data and can be classified into two types—individual-level or group-level fairness. Fairness definition at individual level is based on the premise that similar people should be treated similarly. For example, Dwork et al. (Dwork et al. 2012) defined a task-specific metric based on a probabilistic distance measure between individuals via a Lipschitz condition. The metric is used as constraints to optimize a fairness-aware classifier. In contrast, the group-level approach quantifies fairness in terms of statistical measures such as demographic parity, equalized odds (Hardt et al. 2016) or balanced error rate (Feldman et al. 2015) with respect to the protected groups. The measures are typically computed from a confusion matrix (Berk et al. 2018) and are used to ensure that the average performance do not vary significantly among different groups of a protected attribute.

In addition, there has been growing literature on developing fairness-aware methods. Current methods can be divided into three categories. The first category includes prepossessing algorithms (Zemel et al. 2013; Madras et al. 2018) with the motivation that training data is the main cause of bias in machine learning. Zemel et al. (Zemel et al. 2013) introduced an optimization algorithm to map data points into a new space to ensure membership in the protected group is lost. Madras and et al. (Madras et al. 2018) connected group fairness concept to adversarial concept for learning fair representation. In addition, there has been some recent work on fairness in recommender systems related to the link prediction problem (Zhu, Hu, and Caverlee 2018).

Fairness for Network Data

We first review the fairness criteria for i.i.d. data. Let Y be the target variable of interest (true outcome) and X be a set of input features. Conventional supervised learning algorithms are designed to predict the target outcome Y from X by learning a model f such that $\hat{Y} = f(X)$ is the predicted outcome. Existing fairness-aware methods seeks to ensure

that the predictions generated by the model will not discriminate against one or more subgroups, defined by a protected attribute $X^{(p)}$ such as gender, race, or sexual orientation.

A widely used criterion for assessing fairness is *demographic parity* (Louizos et al. 2015; Kamishima, Akaho, and Sakuma 2011; Johndrow, Lum, and others 2019; Edwards and Storkey 2015; Calders, Kamiran, and Pechenizkiy 2009), which considers the degree of independence between the model output and protected attribute. Assuming both the target outcome and protected attributes are binary-valued, demographic parity seeks to achieve:

$$P(\hat{Y} = 1|X^{(p)}X^{(p)}) = P(\hat{Y} = 1)$$

Another well known fairness criterion is *equalized odds* (Hardt et al. 2016), which seeks to ensure that the predictions are conditionally independent of the protected attribute given the true outcome:

$$P(\hat{Y} = 1|X^{(p)} = 0, Y = y) = P(\hat{Y} = 1|X^{(p)} = 1, Y = y),$$

If we consider $Y = 1$ as advantaged outcome, a special case for this criterion is known as *equal opportunity* (Hardt et al. 2016), which is defined as follows:

$$P(\hat{Y} = 1|X^{(p)} = 0, Y = 1) = P(\hat{Y} = 1|X^{(p)} = 1, Y = 1),$$

Dyadic-level Fairness

In this paper, we investigate the filter bubble problem from the perspective of algorithm fairness. Specifically, a dyadic-level fairness criterion can be defined based on the protected group membership of individuals participating in the links. Below, we consider two such criteria:

- **Subgroup dyadic-level protection**, where fairness is assessed in terms of how representative each protected subgroup is in the formation of the links. For example, in applications such as link-based recommender systems, the fairness criteria could be to ensure that the recommended links do not favor certain subgroups in the population at the expense of other subgroups.
- **Mixed dyadic-level protection**, where fairness is determined based on homogeneity of the nodes involved in each link. Specifically, a link is considered to be an *intra-group link* if it relates a pair of nodes with the same protected attribute values. Otherwise, it is known as an *inter-group link*. To prevent effects such as filter bubble, inter-group or mixed links should be favored to prevent segregation of the users.

In principle, the subgroup dyadic-level protected can be implemented using existing group-level fairness criteria for i.i.d. data by applying them to the links instead of individual nodes in the network. For mixed dyadic-level protected, we introduce the network modularity measure to be described in the next section.

Network Modularity

Homophily (McPherson, Smith-Lovin, and Cook 2001), which is the tendency of individuals to form relations with others similar to them, is an important characteristic of many social networks. Such relationship can be quantified using the well-known network modularity (or assortative mixing) measure (Newman and Girvan 2004; Newman 2006). The measure, which was originally developed for community detection in networks, is based on the idea that a random graph is not expected to contain any clustering structure. Any community structure in a given network can thus be validated by comparing its link density against its expected density if the link structure of the network is completely random. The modularity measure is defined as follows (Newman and Girvan 2004):

$$Q = \frac{1}{2m}\sum_{ij}\Big(A_{ij} - \frac{d_i d_j}{2m}\Big)\delta(c_i, c_j) \tag{1}$$

where A is the adjacency matrix representation of the network, $\delta(c_i, c_j)$ is the Kronecker delta function, c_i is the community of node i, d_i is its corresponding degree and m is total number of links. Intuitively, a network is said to be assortative if a significant portion of its links are between nodes that belong to the same community.

The modularity measure can be used to determine whether a network is unfair in terms of mixed dyadic-level protection by replacing $\delta(c_i, c_j)$ in Equation (1) with $\delta(X_i^{(p)}, X_j^{(p)})$, where $X_i^{(p)}$ is the protected attribute value for node i. The Q value is thus influenced by only those pairs of nodes belonging to the same protected class. Values of Q close to one would indicate high unfairness due to the strong alignment between the link structure and the protected attribute while values close to zero indicate high fairness. For numeric-valued protected attributes such as age or income level, it can be modified as follows:

$$r = \frac{\sum_{ij}(A_{ij} - d_i d_j/2m)X_i^{(p)}X_j^{(p)}}{\sum_{kl}(d_k\delta_{kl} - d_k d_l/2m)X_k^{(p)}X_l^{(p)}}$$

$$\text{where} \qquad \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{o.w} \end{cases}$$

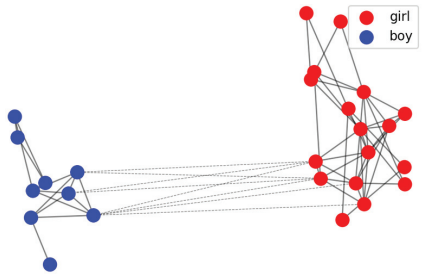
This is also known as assortativity coefficient of the network.

To illustrate the use of modularity as a measure of unfairness, consider the networks shown in Figure 1. The data correspond to friendship relations among freshman at a secondary school in the Netherlands from 2003-2004 (Snijders, Van de Bunt, and Steglich 2010). Using gender as protected attribute, the modularity value for the first network shown in Figure 1(A) is equal to 0.3033 while the value for the second network is 0.0179. Note that the network with higher modularity has more links between students of the same gender compared to the one with lower value, and thus, is unfair from the perspective of mixed dyadic-level protection.

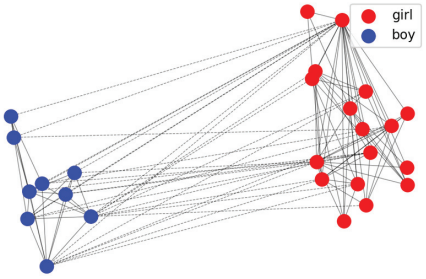
Our proposed fairness-aware framework evaluates the reduction in the modularity measure to determine whether the modified network obtained from the link prediction results is biased towards creating more inter-group or intra-group links. Specifically, we define the following metric:

$$modred = \frac{Q_{\text{ref}} - Q_{\text{pred}}}{Q_{\text{ref}}}, \tag{2}$$

where Q_{ref} is the modularity measure of a reference network (e.g., the ground truth network when evaluating link



(a) Snapshot taken in 2003. Modularity = 0.3033.



(b) Snapshot taken in 2004. Modularity = 0.0179.

Figure 1: Snapshots of friendship relation among students at a Dutch school taken 2003 and 2004 along with their modularity values. The node color represents the student's gender. Darker dashed lines correspond to links between students of different gender while the solid ones correspond to links between students of the same gender.

prediction algorithms) and Q_{pred} is the modularity of the predicted network, i.e., the network obtained by augmenting the predicted links to the original network. A positive $modred$ value indicates that the link prediction algorithm predicts more inter-group links than the ground truth network while a negative value suggests that the algorithm is predicting more intra-group links than the ground truth network.

Greedy Post-Processing

One approach to promoting fairness in link prediction is to post-process the prediction results. To this end, we propose a greedy algorithm for reducing modularity of the predicted network. It takes as input a set of binarized link predictions, $\{\dot{e}_{xy}\}$ and calculates the change in modularity resulting from flipping the prediction of each node pair. The change in modularity for flipping link $\dot{e}_{xy}$ is:

$$score(\dot{e}_{xy}) = \frac{(-1)^{\delta(\dot{e}_{xy})}}{2m}\Big(-1+\frac{d_x+d_y-1}{2m}\Big)\delta(X_x^{(p)},X_y^{(p)}) + \Big(\sum_{\substack{v\in V, X_v^{(p)}\neq X_x^{(p)}\\ v\neq y}} d_v + \sum_{\substack{v\in V, X_v^{(p)}\neq X_y^{(p)}\\ v\neq x}} d_i\Big)/4m^2 \tag{3}$$

where the value of $(-1)^{\delta(\dot{e}_{xy})}$ is -1 if $\dot{e}_{xy}$ is 1 and $+1$ otherwise, d_x and d_y are the degrees of nodes x and y respectively. After computing this score for each predicted link we flip the edges with the lowest scores. This is another approximation since the score for edge should be recomputed after each edge is flipped due to changes in the value of d_x and d_y. The number of link predictions to flip is a hyper-parameter that can be varied depending on the importance of accuracy versus modularity.

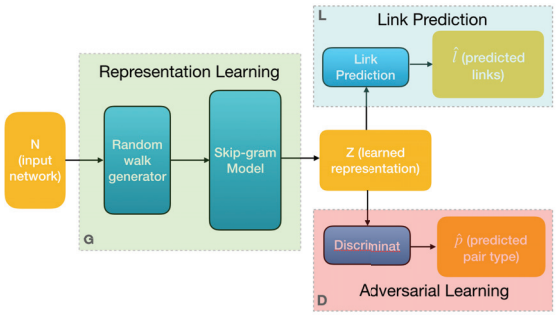


Figure 2: FLIP architecture

Adversarial Learning for Fair Link Prediction

Consider an attributed network $\mathcal{N} = (V, E, X)$, where V is the set of nodes, $E \subseteq V \times V$ is the set of links, and $X \in \mathbb{R}^{|V|\times d}$ is a matrix corresponding to the set of attribute values associated with the nodes in V. Assume X can be partitioned into submatrices $[X^{(p)}, X^{(u)}]$, which correspond to the protected and unprotected features of the nodes. Here we only consider a binary value $X^{(p)}$. Our goal is to accurately infer new links in the network without being biased against the formation of inter-group links.

Our proposed framework, known as FLIP (Fairness-aware LInk Prediction), employs an adversarial learning approach to ensure that inter-group links are well-represented among the predicted links. The framework consists of the following 3 components, as illustrated in Figure 2:

1. A generator, G, that takes the attributed network as input and learns a representation $G(v)$ for each node $v \in V$. We use DeepWalk (Perozzi, Al-Rfou, and Skiena) as the generator, though in principle, the framework can be applied to other network representation learning methods.
2. A discriminator, D, that takes the representations for each pair of nodes produced by the generator as input and attempts to predict if it is an intra-group or inter-group node pair. The discriminator's predicted probability that a pair of nodes has the same protected attribute value is denoted as $D(G(v_i), G(v_j))$
3. A link prediction component, L, which tries to infer new links given node representation learned by the generator. The predicted probability that a link exists between a pair of vertices is $L(G(v_i), G(v_j))$.

To understand the rationale behind the framework, note that a good feature representation learned by the generator will enable the link prediction component to infer correctly whether a node pair is connected. If the link structure of the network is biased towards intra-group links, so will the link

prediction component as well as the generator. The discriminator plays the role of an adversary who attempts to predict whether a node pair involves nodes from the same group or from different groups. By making the generator and discriminator to work against each other, this would lead to a situation in which the generator produces a feature representation that is good enough for link prediction yet unbiased enough to prevent the discriminator from inferring whether it is an inter-group or intra-group node pair. In networks with homophily property, this will help to discourage the prediction of intra-group links and promotes more inter-group links.

Discriminator

In recent years, adversarial networks have been used to achieve different fairness criteria for independent and identically distributed (i.i.d) data (Beutel et al. 2017; Madras et al. 2018). The shared idea between these methods is an adversarial component that attempts to predict the protected attribute value X_u. A naïve approach to achieving fairness in network data is to follow same path and design an adversarial component that predicts the protected attribute of a node using the following cross entropy cost function:

$$J^D = -\frac{1}{|V|}\sum_{v\in V}\Big[X_v^{(p)} log(\hat{y}_v) + (1 - X_v^{(p)})\log(1-\hat{y}_v)\Big]$$

Here $\hat{y}_u = D(G(u))$ is the prediction of the discriminator of the binary protected value of node u.

However this will not necessarily result in mixed dyadic level protection because intra-group links may still be favored in a homophilic network. To solve this challenge we propose the following adversarial loss:

$$J^D = -\frac{1}{|\mathcal{T}|}\sum_{(u,v)\in\mathcal{T}}\Big[p_{uv} log(\hat{p}_{uv}) + (1-p_{uv})\log(1-\hat{p}_{uv})\Big] \quad (4)$$

where $\mathcal{T} \subseteq V \times V$ is the set of node pairs in the training data, p_{uv} is the actual type of node pair (u, v) with respect to a given protected attribute (i.e. intra-group vs inter-group) and $\hat{p}_{uv}$ is the discriminator's prediction. Instead of inferring the node's protected attribute, the discriminator receives a pair of node representations, which it passes to a two layer fully connected network with leaky ReLU activation to determine the probability that it is an intra-group node pair.

Generator

In contrast to the original GAN framework proposed by (Goodfellow et al. 2014) where the generator seeks to generate samples of data points that seem real, the generator in our framework tries to learn node representation that will preserve important structural information of the network without implicit usage of the protected attribute information. For the generator, we utilized DeepWalk (Perozzi, Al-Rfou, and Skiena) which is a network representation learning method inspired by the Skip-gram (Mikolov et al. 2013) model from natural language processing. DeepWalk consists of two steps: the first step is to extract sequences of nodes from the network by performing a series of truncated random walks starting from each node in the input network. In the second step, the node sequences generated from the random walk process are used to learn the feature representation of each node. This is accomplished as follows. A sliding window of width w scans the generated node sequences to generate all the node pairs (u, v) in which node v appears in the sliding window centered at node u. A fully connected neural network with a single hidden layer predicts the probability of the occurrence of node v given the one hot encoding, $\bar{u}$, of node u. Specifically, the network attempts to predict $p(v|u)$ for each u as follows:

$$\mathbf{p}(v|u) \simeq \frac{exp(f'(v)^\top f(u))}{\sum_{v'\in V} exp(f'(v')^T f(u))} \quad (5)$$

where $f(v) = W\bar{v}$, $f'(u) = Z\bar{u}$, W is the weight matrix between the input and hidden layers of the network, and Z is the weight matrix between the hidden and output layers of the network. The rows of matrix W are the node representations generated by the skip-gram model so we have $G(u) = f(u)$.

The parameters of DeepWalk are trained using the maximum likelihood estimation approach, with the following loss function:

$$J^{Skip} = -\sum_{u\in V}\Big[-log(\sum_{v'\in V} exp(f'(v')^\top f(u)) + \sum_{v'\in\Omega_w(u)} exp(f'(v)^\top f(u))\Big]$$

Here $\Omega_w(u)$ represent the set of all nodes that appears in the neighborhood of node u in the given random walk sequence with window size of width w.

Link prediction

This component takes a pair of node embeddings as input to predict whether their nodes should be linked or not. This is accomplished by adding a two-layer link prediction network to the GAN model. During the training phase the link prediction component receives pairs of node embeddings and concatenates them into a feature vector, which is then passed to a two-layer fully connected network with leaky ReLU activation. The output of the network corresponds to the likelihood of a link to exist between the node pair. Here we deployed the standard cross entropy cost function as follows:

$$J^L = -\frac{1}{|\mathcal{T}|}\sum_{(u,v)\in\mathcal{T}}\Big[e_{uv} log(\hat{e}_{uv}) + (1-e_{uv})\log(1-\hat{e}_{uv})\Big] \quad (6)$$

where $\hat{e}_{uv}$ is output of the link prediction component for node pair (u, v) and e_{uv} is the binary ground truth link label.

Putting everything together, the overall loss function for the proposed framework is given as follows:

$$J^G = (1-\alpha)J^{Skip} - \alpha J^D + \beta J^L \quad (7)$$

where β is a hyperparameter. The generator, discriminator, and link prediction network are all trained end to end using Adam (Kingma and Ba 2014). The generator and link

Table 1: data sets

network	#nodes	#edges	protected feature
Dutch school	26	221	gender
Facebook	1,034	26,749	gender
Google+	4,938	547,923	gender

Table 2: Proximity based link prediction algorithms. For each node v, $N(v)$ is the set of its immediate neighbors.

Method	Definition
Jaccard	$\frac{\lvert N(u)\cap N(v)\rvert}{\lvert N(u)\cup N(v)\rvert}$
Adamic/Adar	$\sum_{x\in N(u)\cap N(v)} \frac{1}{log(d_x)}$
Preferential attachment	$d_x d_y$

	$p(link)$	$p(link\|intra)$	$prob(link\|inter)$
Dutch School	0.3662	0.5406	0.1699
Facebook	0.050	0.053	0.047
Google+	0.045	0.061	0.0316

Table 3: Homophily effect

prediction network are trained on the same batches but every other batch is used to train the discriminator only so that training alternates between updating the link predictor and generator together on one batch and updating the discriminator on the next batch.

Experimental Evaluation

This section describes the experiments performed to evaluate the efficacy of our proposed methods to address the filter bubble problem in network link prediction.

Experiment Setup

We first discuss the experimental setup, including data sets, baselines and evaluation metrics used in our experiment.

Datasets We evaluated our methods on three real world social network data sets. Table 1 summarizes the main properties of these data sets. The first data set is a Facebook ego network (Leskovec and Mcauley 2012), which contains 1,034 nodes, and 26,749 friendship links. The second data set is Google+, which has 4,938 nodes and more than 500,000 links. (Leskovec and Mcauley 2012), The third data set is Dutch school network (Snijders, Van de Bunt, and Steglich 2010), which corresponds to friendship relations among 26 freshmen at a secondary school in the Netherlands. For all three datasets, we use gender as the protected attribute for inferring intra-group and inter-group links.

Baseline Algorithms We considered 4 state-of-art link prediction algorithms as baselines. Three of them are well known classical proximity based methods which use neighborhoods structural information. The first baseline is based on the well known Jaccard's (Jac) coefficient similarity metric which is deployed in the context of network link prediction by calculating the portion of common neighbors for a given nodes pair. The second baseline is Adamic/Adar(Ad-Ad), a similar measure that assigns less weight to more connected common neighbors. The third proximity based algorithm is preferential attachment (Pr-At)(Mitzenmacher 2001) which sets the probability of a connection between two pair of such that it is correlated with the product of the their degrees (Newman 2001). Table 2 summarized the formal definition of these algorithms. We also considered the more recent DeepWalk(DW) algorithm (Perozzi, Al-Rfou, and Skiena) which learns d-dimensional feature representations of nodes by simulating uniform random walks and provides latent features for nodes at the first step and then, similar to proposed approach in (Grover and Leskovec), we construct the edge embedding by applying binary Hadammard product operation to the given node pair and train a logistic regression to do link prediction. For evaluation, we use the settings suggested in the original DW paper for both the DW baseline and the proposed method's skip gram model. These settings are: latent feature dimension (128), length of random walks(80), and number of random walks(10) and window size(10) on all data sets.

We also consider a traditional fairness algorithm based on equalized odds which we use to post-process our 4 baselines. As previously mentioned, imposing an equalized odds constraint on the predictions of a model is a popular way of ensuring fair predictions. For our task, we use a generalized version of equalized odds proposed in (Pleiss et al. 2017) to post-process each of the baseline algorithm predictions. To make the generalized equalized odds constraint compatible to network data setting we treat link type, intra-group versus inter-group, as each link's binary protected attribute. We refer to this post processing algorithm as (PEO).

Sampling process and training A big challenge for link prediction algorithms is the sparsity of real world network data. In other words, since the number of existing links are significantly smaller than non-existing links, training a model which is not biased toward negative examples is difficult. Given a graph $N = (V, E, X)$ we generate a training set with equal number of negative and positive examples $< N', E^+, E^- >$. Here N' is the remaining sub-graph after removing all sampled positive links, E^+, and E^- is a set of randomly sampled non-links such that $|E^+| = |E^-|$. Sampling positive links from N is random with the restriction that N' remains connected. For each data set we generating 10 examples of $< N', E^+, E^- >$ by deleting 80% of all links in N. For FLIP and DW we learn node representations by performing random walks on graph N' and train the link prediction using 10% of the generated positive and negative samples. For the other baselines we used all the 30% of E^+ and E^- for calculating the proximity measures. We used remaining 70% for test.

Evaluation Metric We evaluate the quality of link predictions with two metrics, accuracy and the area under the ROC curve (AUC) which represents the trade-off between true and false positives with respect to different thresholds. For the fairness subgroup dyadic-level metric we consider modred measure given in equation 2.

Table 4: Performance comparison between baseline and proposed algorithms on 3 real-world datasets. Results are reported based on the average AUC and modred scores after repeating the sampling process 10 times.

Method	Dutch school		Facebook		Google+	
	AUC	$modred$	AUC	$modred$	AUC	$modred$
Jac	0.6500 +/- 0.0008	-0.5030 +/- 0.0046	0.8305 +/- 0.0	-0.1494 +/- 0.0396	0.7932 +/- 0.0	0.0932 +/- 0.0297
Ad-Ad	0.6571 +/- 0.0006	-0.3761 +/- 0.0044	0.836 +/- 0.0	0.2224 +/- 0.0089	0.8692 +/- 0.0	-0.3048 +/- 0.0015
Pr-At	0.6023 +/- 0.0016	0.4431 +/- 0.0022	0.8068 +/- 0.0	0.4601 +/- 0.0015	0.9047 +/- 0.0	-2.9354 +/- 0.0106
DW	0.5287 +/- 0.0074	-0.0471 +/- 0.2423	0.951 +/- 0.0	0.0889 +/- 0.0022	0.7708 +/- 0.0006	0.1663 +/- 0.0386
Jac+PEO	0.5356 +/- 0.003	0.0325 +/- 0.0521	0.7992 +/- 0.0001	-0.5575 +/- 0.2353	0.7500 +/- 0.0	3.4696 +/- 0.048
Ad-Ad+PEO	0.5275 +/- 0.0024	-0.2337 +/- 0.052	0.7992 +/- 0.0001	0.0132 +/- 0.0599	0.8292 +/- 0.0	3.2193 +/- 0.0133
Pr-At+PEO	0.5054 +/- 0.0003	0.0219 +/- 0.1091	0.6822 +/- 0.0004	-0.208 +/- 0.2277	0.8584 +/- 0.0	3.8539 +/- 0.1538
DW+PEO	0.4908 +/- 0.0055	-0.1209 +/- 0.2133	0.9489 +/- 0.0	0.0142 +/- 0.0204	0.7354 +/- 0.0008	3.546 +/- 2.6518
Jac+GM	0.6571 +/- 0.0289	-0.2179 +/- 0.0840	0.8421 +/- 0.0018	0.6613 +/- 0.0501	0.7399 +/- 0.0013	0.9657 +/- 0.0405
Ad-Ad+GM	0.6528 +/- 0.0265	-0.2110 +/- 0.0882	0.8421 +/- 0.0018	0.6613 +/- 0.0501	0.8179 +/- 0.0009	1.3693 +/- 0.0383
Pr-At+GM	0.5827 +/- 0.0440	0.1795 +/- 0.1654	0.7400 +/- 0.0036	0.9190 +/ 0.0597	0.8422 +/- 0.0004	1.7478 +/- 0.0960
DW+GM	0.5363 +/- 0.0560	0.1335 +/- 0.0907	0.9013 +/- 0.0045	0.4972 +/- 0.0617	0.7254 +/- 0.0295	1.5062 +/- 0.5906
FLIP	0.6576 +/- 0.0039	0.3592 +/- 0.0089	0.8601 +/- 0.0001	0.3483 +/- 0.0039	0.8575 +/- 0.0	0.2071 +/- 0.0088

Experimental Results

In the following subsection we investigate the general performance of the proposed framework.

Homophily property

Table 3 summarizes our evaluation on the homophily property of the three networks. The first column is the probability a node pair is linked. The second column shows the conditional probability a node pair is linked given that it is an intra-group node pair, while the third column corresponds to the conditional probability of a link between an inter-group node pair. These probabilities indicate that all three networks have homophily property because they are more likely to have intra-group links than inter-groups links.

Performance Comparison We summarize our results for link prediction in Table 4. For FLIP we report the result for $\alpha = 0.1$ and $\beta = 0.2$. For greedy post-processing we chose to invert 3% of the predictions that reduces the modularity the most. Based on these results we can make several observations. First, there is generally a trade off between AUC and $modred$ so higher $modred$ scores are only achievable by sacrificing accuracy.

Second, none of the baseline algorithms achieve consistently high $modred$ scores. In particular, equalized odds post-processing provides highly inconsistent gains in $modred$ that are heavily dependant on the data set. It provides significant gains on the Google+ data set, but on the Dutch school data set it provides only moderate gains. It is also a moderate impediment on the Facebook data set. However, greedy post-processing and FLIP always achieve high $modred$ scores and provide a good balance between AUC and $modred$. This is unsurprising since FLIP and greedy post-processing were the only two techniques specifically designed for promoting fairness in link prediction. This demonstrates the importance of developing algorithms tailored specifically for network data and link prediction.

Third, among the baselines, preferential attachment does the best job in terms of balancing the tradeoff between accuracy and $modred$. One possible explanation for this is that all other baselines make predictions based on the neighborhood structure of nodes. In a network that is homophillic with respect to a protected attribute, nodes with the same protected attribute value are likely to have similar neighborhood structure. Since all of our networks are homophillic with respect to the protected attribute, link prediction methods based on neighborhood structure are more likely to reinforce the existing homophilly and create intra-group links. In contrast, preferential attachment ignores the neighborhood structure of nodes when making predictions so it less affected by pre-existing network homophilly.

Conclusions

This paper presents novel fairness-aware methods to alleviate the filter bubble problem in network link prediction. First, we present a fairness criterion based on network modularity measure to determine whether inter-group links are well-represented in the predicted output of a link prediction algorithm. We then consider two approaches to overcome the filter bubble problem—one based on a greedy postprocessing approach using the *modred* measure while the other based on an adversarial learning framework. Experimental results showed that the proposed methods are promising as they can reduce modularity of the predicted network without degrading prediction accuracy significantly.

References

Al Hasan, M.; Chaoji, V.; Salem, S.; and Zaki, M. 2006. Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security*.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias. *ProPublica, May* 23.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*.

Beutel, A.; Chen, J.; Zhao, Z.; and Chi, E. H. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*.

Clauset, A.; Moore, C.; and Newman, M. E. 2008. Hierarchical structure and the prediction of missing links in networks. *arXiv preprint arXiv:0811.0484*.

Crandall, D. J.; Backstrom, L.; Cosley, D.; Suri, S.; Huttenlocher, D.; and Kleinberg, J. 2010. Inferring social ties from geographic coincidences. *PNAS* 107(52):22436–22441.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*.

Edwards, H., and Storkey, A. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proc. of KDD*. ACM.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*, 2672–2680.

Grover, A., and Leskovec, J. node2vec: Scalable feature learning for networks. In *Proc. of KDD*.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *NeurIPS*, 3315–3323.

Hofstra, B.; Corten, R.; Van Tubergen, F.; and Ellison, N. B. 2017. Sources of segregation in social networks: A novel approach using facebook. *American Sociological Review* 82(3):625–656.

Johndrow, J. E.; Lum, K.; et al. 2019. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics* 13(1):189–220.

Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness-aware learning through regularization approach. In *ICDM Workshops*, 643–650. IEEE.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *NeurIPS*, 4066–4076.

Leskovec, J., and Mcauley, J. J. 2012. Learning to discover social circles in ego networks. In *NeurIPS*, 539–547.

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. Gated graph sequence neural networks.

Li, X.; Du, N.; Li, H.; Li, K.; Gao, J.; and Zhang, A. 2014. A deep learning approach to link prediction in dynamic networks. In *SDM*. SIAM.

Liben-Nowell, D., and Kleinberg, J. 2007. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology* 58(7):1019–1031.

Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*.

Masrour, F.; Barjesteh, I.; Forsati, R.; Esfahanian, A.-H.; and Radha, H. 2015. Network completion with node similarity: A matrix completion approach with provable guarantees. In *ASONAM*, 302–307. IEEE.

Masrour, F.; Tan, P.-N.; Esfahanian, A.-H.; and VanDam, C. 2018. Attributed network representation learning approaches for link prediction. In *ASONAM*, 560–563. IEEE.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27(1):415–444.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mitzenmacher, M. 2001. A brief history of lognormal and power law distributions. In *Proceedings of the Allerton conference on communication, control, and computing*.

Newman, M. E., and Girvan, M. 2004. Finding and evaluating community structure in networks. *Physical review E* 69(2):026113.

Newman, M. E. 2001. Clustering and preferential attachment in growing networks. *Physical review E* 64(2):025102.

Newman, M. E. 2006. Modularity and community structure in networks. *PNAS* 103(23):8577–8582.

Nguyen, T. T.; Hui, P.-M.; Harper, F. M.; Terveen, L.; and Konstan, J. A. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *WWW*.

O'Neil, C. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pariser, E. 2012. The filter bubble: What the internet is hiding.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. Deepwalk: Online learning of social representations. In *Proc. of KDD*.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On fairness and calibration. In *NeurIPS*.

Scripps, J.; Tan, P.-N.; Chen, F.; and Esfahanian, A.-H. 2008. A matrix alignment approach for link prediction. In *Proc of ICPR*.

Snijders, T. A.; Van de Bunt, G. G.; and Steglich, C. E. 2010. Introduction to stochastic actor-based models for network dynamics. *Social networks* 32(1):44–60.

Tian, F.; Gao, B.; Cui, Q.; Chen, E.; and Liu, T.-Y. 2014. Learning deep representations for graph clustering. In *AAAI*.

Wang, D.; Pedreschi, D.; Song, C.; Giannotti, F.; and Barabasi, A.-L. Human mobility, social ties, and link prediction. In *Proc. of KDD*.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *ICML*, 325–333.

Zhu, Z.; Hu, X.; and Caverlee, J. 2018. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM.