The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Fairness in Network Representation by Latent Structural Heterogeneity in Observational Data

Xin Du, Yulong Pei, Wouter Duivesteijn, Mykola Pechenizkiy
Technische Universiteit Eindhoven
the Netherlands
{x.du, y.pei.1, w.duivesteijn, m.pechenizkiy}@tue.nl

Abstract

While recent advances in machine learning put many focuses on fairness of algorithmic decision making, topics about fairness of representation, especially fairness of network representation, are still underexplored. Network representation learning learns a function mapping nodes to low-dimensional vectors. Structural properties, e.g. communities and roles, are preserved in the latent embedding space. In this paper, we argue that latent structural heterogeneity in the observational data could bias the classical network representation model. The unknown heterogeneous distribution across subgroups raises new challenges for fairness in machine learning. Pre-defined groups with sensitive attributes cannot properly tackle the potential unfairness of network representation. We propose a method which can automatically discover subgroups which are unfairly treated by the network representation model. The fairness measure we propose can evaluate complex targets with multi-degree interactions. We conduct randomly controlled experiments on synthetic datasets and verify our methods on real-world datasets. Both quantitative and quantitative results show that our method is effective to recover the fairness of network representations. Our research draws insight on how structural heterogeneity across subgroups restricted by attributes would affect the fairness of network representation learning.

1 Introduction

There are increasing demands for machine learning on diverse real-world applications such as policing (Brennan, Dieterich, and Ehret 2009), lending (Mahoney and Mohen 2007) and credit scoring (Khandani, Kim, and Lo 2010). Fair decision making has become more and more important for machine learning research. Several notions have been defined for algorithmic fairness (Dwork et al. 2012; Hardt et al. 2016; Zafar et al. 2015). Among these methods, fairness is measured for individuals or pre-defined groups based on statistical quantities like false positive / negative rates or classification rates. Recently, more and more papers notice that the fairness of decision making process is highly dependent on biases which already exist in data collection process (Chen, Johansson, and Sontag 2018). In parallel, fairness of representation learning receives a lot of attentions (Edwards and Storkey 2015; Song et al. 2018; Madras et al. 2018). Among these methods, people are trying to learn similar representations for different groups, to ensure that the consequent decision making is independent of group attributes (Zhao and Gordon 2019).

Despite the recent research focus on fair machine learning, the study of fair representation in networks still lacks exploration. Comparing with existing work, the challenges are two-fold: on the one hand, unlike statistical quantities of single decision variables, fairness of network representation requires to compare multi-degree interactions between nodes. We need to develop a new statistical measure to evaluate the differences between node representations. On the other hand, as pointed out by some research, when we only ensure fairness for some small amount of pre-defined subgroups, it might actually *increase* rather than decrease model discrimination (Kearns et al. 2017). In order to prevent this problem, we propose to investigate the fairness of network representation by generating subgroups with regard to any combinations of attributes. Computational cost would be very high due to the exponentially increasing amount of subgroups. We tackle this problem by employing *exceptional model mining* (Duivesteijn, Feelders, and Knobbe 2016), a framework of generating and evaluating subgroups by heuristically exploring the attribute space.

Before discussing fairness of network representation, we firstly focus on *structural heterogeneity* in networks. Unknown heterogeneity across the data can lead a model to be very effective for some subpopulations and ineffective for some other subpopulations (Pearl 2017). We argue in this paper that the potential unfairness of network representation is associated with the structural heterogeneity in networks. In Figure 1, we demonstrate a toy example of structural heterogeneity and show how it can affect the network representation. As we can see, the network structure in subgroups '$x = 1$' and '$x = 0$' are very different from each other. A random walk based neighborhood function will generate different distributions of nodes in neighborhoods conditioned on different attributes. The classical network representation model could be biased. These biased representations might lead to unfairness of consequential decision making models.

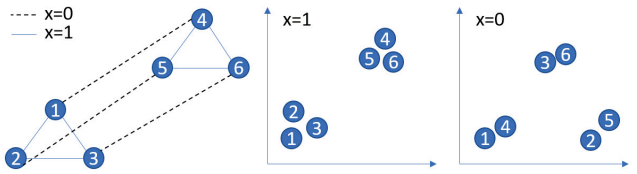


Figure 1: Toy example: dashed lines represent edges with attribute $x = 0$, solid lines represent edges with attribute $x = 1$. Obviously, the distributions of nodes in neighborhoods conditioned on different attributes ($P(N(V_o)|x = 1)$, $P(N(V_o)|x = 0)$) are different. This can lead to very different representation functions.

The study of fair machine learning should prevent the propagation of bias from the data to modeling results (Madras et al. 2019).

In this paper, we propose to analyze the latent structural heterogeneity across subgroups and discuss its effects on the fairness of network representations. Top-Q subgroups with highest measurement scores are reported to recover the fairness of a network representation model. In order to investigate whether the reported subgroups represent significant signals in the data, we conduct hypothesis testing against random noise.

1.1 Main Contributions

- We study the problem of fairness in terms of the latent structural heterogeneity across subgroups in networks. As far as we know, this is the first work which considers structural heterogeneity to measure the fairness of network representation.
- We propose a new measurement, mean latent similarity discrepancy (MLSD) to quantify the differences between node representations. MLSD can calculate the statistical discrepancy between node representations which is sensitive to structural heterogeneity.
- We conduct hypothesis testing to verify the significance of fairness score, distinguishing structural discrepancy from randomized noise. We design a series of randomized experiments on synthetic and real-world datasets to evaluate our method qualitatively and quantitatively.

2 Related Work

Previous work on fair machine learning mainly focuses on the level of a group or individual. Pre-defined sensitive attributes are required, which is not applicable in many real-world applications (Kearns et al. 2017). Fairness on groups is normally measured by statistical parity, which requires positive / negative rate to be equal across groups with regard to sensitive variables (Hardt et al. 2016). Fairness on individuals requires similar individuals to be treated similarly by the models (Dwork et al. 2012). In contrast, fairness of network representation requires to compare more complex relations rather than a single decision variable. For this reason, we propose MLSD which focuses on measuring the statistical discrepancy between node representations.

Representation learning is specified to learn multiple degrees of similarities between units (Mikolov et al. 2013b) in large datasets. This technique is widely used to discover word similarities known as word embedding (Mikolov et al. 2013a) and node similarities known as graph embedding (Hamilton, Ying, and Leskovec 2017). Network representation learning enables us to learn low-dimensional vector representations for nodes from their neighborhood structures. There is a lot of work on learning vector representations of nodes in graphs (Perozzi, Al-Rfou, and Skiena 2014; Grover and Leskovec 2016) . Most existing work on fairness of representation focuses on adversely learning fair representations across groups and preserving highly predictive information for decision making (Zemel et al. 2013). Conversely, we focus on fairness of network representation, which requires definition of a new measurement with regard to the structural heterogeneity in networks. Our work can help people understand how structural heterogeneity is correlated with attributes and how unfairness of network representation exists by heuristically discovering subgroups.

The aim of Subgroup Discovery (SD) (Klösgen 1996; Wrobel 1997; Herrera et al. 2011; Atzmueller 2015) is to find subsets restricted by descriptive attributes, in which the distribution of one predefined target variable is substantially different from the distribution in the whole dataset. Exceptional Model Mining (EMM) (Duivesteijn, Feelders, and Knobbe 2016) can be seen as an extension of SD, which focuses on multiple target variables. EMM can integrate various model classes with different performance measures to adapt to different tasks, e.g. Bayesian networks (Duivesteijn et al. 2010). Most of the existing model classes cannot handle structural properties in networks. Weighted relative accuracy was introduced to evaluate characteristics in subgraph (Bendimerad, Plantevit, and Robardet 2016), first-order Markov chains have been introduced as a model class for sequential data (Lemmerich et al. 2016). However, structural properties, especially role structures (Jin, Lee, and Hong 2011) are not considered in those methods.

In EMM, a quality measure is defined to evaluate the differences between the target models within and outside of the subgroup. Popular examples of quality measures include WRAcc (Todorovski, Flach, and Lavrač 2000) and z-score (Mampaey et al. 2015). In order to compare the network representations which preserve the structural properties, we design the Mean Latent Similarity Discrepancy (MLSD) quality measure, based on the U-statistic (Korolyuk and Borovskich 2013). MLSD calculates the mean discrepancy between latent similarities of node vectors, reflecting the statistical difference between network representations.

3 Problem Setup

We assume a dataset Ω: a set of M nodes $v \in V$ and a bag of N records $\mathbf{r} \in \Omega$ of the form $\mathbf{r} = (x_1, \ldots, x_k, v_o, v_d)$, where k is a positive integer and v_o, v_d refer to a directed edge from the origin v_o to the destination v_d (cf. Table 1). We call $x_1, \ldots, x_k$ descriptive variables, and v_o, v_d target variables. The descriptive variables are taken from an unrestricted domain $\mathcal{A}$. Mathematically, we define descriptions

Records	Descriptive Variables	Target Variables
r^1	$x_1^1, \ldots, x_k^1$	v_o^1, v_d^1
$\vdots$	$\vdots \quad \ddots \quad \vdots$	$\vdots$
r^n	$x_1^n, \ldots, x_k^n$	v_o^n, v_d^n

Table 1: A network dataset of N edges over a set of nodes $V = \{v_1, \ldots, v_m\}$ and attributes X= $\{x_1, \ldots, x_k\}$.

as functions $D : \mathcal{A} \rightarrow \{0, 1\}$. A description D covers a record $\mathbf{r}^i$ if and only if $D(x_1^i, \ldots, x_k^i) = 1$.

Definition 1 (Subgroup) *A subgroup corresponding to a description D is the bag of records $S_D \subseteq \Omega$ that D covers, i.e.*

$$S_D = \left\{ \boldsymbol{r}^i \in \Omega \middle| D(x_1^i, \ldots, x_k^i) = 1 \right\}.$$

Definition 2 (Quality Measure) *A quality measure is a function $\varphi : \mathcal{D} \rightarrow \mathbb{R}$ that assigns a numeric value to a description D. Occasionally, we use $\varphi(S)$ to refer to the quality of the induced subgroup: $\varphi(S_D) = \varphi(D)$.*

Typically, a quality measure assesses the subgroup at hand based on the target variables. Hence, a description and a quality measure interact through different partitions of the dataset columns; the former focuses on the descriptors, the latter focuses on the targets, and they are linked through the subgroup.

We can model the network as $G_D = (V, E, X, D)$, where V represents set of M nodes, E set of N edges, X attributes attached on E, and D a description which is satisfied by X. We can define the neighborhood $N(v_o) \subset V$ as a set of nodes generated by a sampling strategy starting from node v_o. In this paper, we consider local community structures, though our method can be easily extended to global role structure (Ribeiro, Saverese, and Figueiredo 2017; Pei et al. 2018). By defining the neighborhood function, we can formulate a distribution of nodes in neighborhoods conditioned on attributes $P(N(v_o)|D)$. If there is structural heterogeneity in networks, then we could have $P(N(v_o)|D) \neq P(N(v_o))$, and $P(N(v_o)|D_1) \neq P(N(v_o)|D_2)$ when $D_1 \neq D_2$. We would use this property to build the measurement for fairness of network representation.

By following Skip model (Mikolov et al. 2013b), we can learn a function $\theta : V \rightarrow \mathbb{R}^l$, which maps each node $v \in V$ to a l-dimensional vector representation. We select θ_D to maximize the probability of visiting neighborhoods $N_D(v_o)$ for each node in network: $\theta_D = \text{argmax}_{\theta_D} \prod_{v_o \in V} p(N_D(v_o)|\theta_D(v_o))$, where $\theta_D(v_o)$ can be represented as u_o. We can formulate the problem of fairness in network representation as an optimization problem of searching subgroups with highest quality scores:

Problem 1 *Given a dataset Ω and a quality measure φ, our task is to find a sequence of Q descriptions $h = \{D_1, \ldots, D_Q\}$, such that $\forall D' \in \mathcal{D} \setminus h$, $\varphi(D') < \varphi(D)$, $\forall D \in h$.*

4 Methodology

Node representations preserve the structural properties from the original networks. In order to measure the fairness across subgroups, we would like to evaluate the difference between node representations learned from that subgroup and learned from the whole dataset. To realize that, at first we need to elicit a latent similarity matrix Z_D, which indicates the similarities between each node and any other nodes:

$$Z_D^{ij} = \frac{d(u_i, u_j)}{\sum_{j \neq i}^{V} d(u_i, u_j)},$$

where $d(u_i, u_j)$ is a distance measure between node i and j in the latent embedding space, and $\sum_{j \neq i}^{V} d(u_i, u_j)$ is a normalizer that ensures $\sum_{j \neq i}^{V} Z_D^{ij} = 1$. Note that we do not consider self loop edges so we let $d(u_i, u_i) = 0$. Now we can compare the latent similarity matrix Z_D from candidate subgroup with Z_Ω from the whole data by using U-statistics (Korolyuk and Borovskich 2013):

$$\varphi_u(D) = \frac{1}{m(m-1)} \sum_{i=0}^{m} \sum_{j \neq i}^{m} |Z_D^{ij} - Z_\Omega^{ij}|_1.$$

By virtue of variance, heterogeneous structures are likely to occur in small subsets of the dataset (Duivesteijn, Feelders, and Knobbe 2016), which are not the results we want. To combat this problem, we incorporate the size of subgroups in the quality measure, by considering the entropy of the split between the records in subgroups and the rest of the records (Duivesteijn et al. 2010):

$$\varphi_{\text{ent}}(D) = -\frac{|D|}{n} \log_2 \left(\frac{|D|}{n} \right) - \frac{n - |D|}{n} \log_2 \left(\frac{n - |D|}{n} \right).$$

The final quality measure can be derived as:

$$\varphi_{\text{MLSD}}(D) = \sqrt{\varphi_{\text{ent}}(D)} \cdot \varphi_u(D).$$

By this quality measure, higher $\varphi_{\text{MLSD}}(D)$ indicates more unfair the network representation is on that subgroup. By applying a search method guided by $\varphi_{\text{MLSD}}(D)$, we can derive the solution for problem 1.

4.1 Statistical Test

In Problem 1, we report the top-Q subgroups with most highest scores calculated by quality measure. However, we do not know whether the scores are significant enough or just a bit of differences because of the random noises. To solve this problem, we assume that the reported vector of top-Q scores is a random draw from distribution P. We propose to independently run our method several times to generate a set of samples from P, denoted by $\mathbf{H} := \{h_1, \cdots, h_x\}$. On the other hand, we randomly shuffle the original data, by permuting the attribute vectors attached with edges in row (Batagelj and Brandes 2005). This can break the dependencies between descriptive variables and targets, and build datasets where the descriptive variables are independent with network structures. After that, we apply our method on each of the shuffled datasets to generate false discoveries[1]. By doing this, we can generate a set of samples from

[1]Because now we already know the ground truth: the descriptive variables and network structures are independent.

the distribution of false discoveries (P_{DFD}) (Duivesteijn and Knobbe 2011), denoted by $\tilde{\mathbf{H}} := \{\tilde{h_1}, \cdots, \tilde{h_y}\}$. Now we can build the null hypothesis by assuming that $\mathbf{H}$ and $\tilde{\mathbf{H}}$ are from the same distribution:

Hypothesis 1 *$\mathbf{H}$ and $\tilde{\mathbf{H}}$ come from the same distribution.*

If the null hypothesis is rejected, then we can be confident that the top-Q subgroups reported by our method are statistically significant. We can define the problem as:

Problem 2 *Let h and $\tilde{h}$ be random variables defined on a topological space $\mathcal{H}$, with distribution P and P_{DFD}. $\mathbf{H} := \{h_1, \cdots, h_x\}$ and $\tilde{\mathbf{H}} := \{\tilde{h_1}, \cdots, \tilde{h_y}\}$ are defined as independently and identically distributed samples from P and P_{DFD} respectively. The problem is to establish a statistical test and conduct hypothesis testing to decide whether $P = P_{DFD}$.*

The main challenge for Problem 2 is that h and $\tilde{h}$ are multivariate (Q-length) and we do not have any prior knowledge about distribution P and P_{DFD}. Hence, classic Student's t-test and Hotelling's T^2-test are not appropriate. Inspired by (Gretton et al. 2012), we use an integral probability metric (Müller 1997) based on distances between Hilbert space mean embeddings of probability distributions, termed as maximum mean discrepancy (MMD). Let $\mathcal{F}$ be a family of functions $f : \mathcal{H} \to \mathbb{R}$, we have:

$$MMD[\mathcal{F}, P, P_{DFD}] := \sup_{f \in \mathcal{F}} (\mathbb{E}_P[f(h)] - \mathbb{E}_{P_{DFD}}[f(\tilde{h})]),$$

where h and $\tilde{h}$, P and P_{DFD} follow Problem 2. Empirically, we can derive the unbiased estimate of the squared MMD in terms of kernel functions ψ as:

$$MMD_u^2[\mathcal{F}, \mathbf{H}, \tilde{\mathbf{H}}] = \frac{1}{x(x-1)} \sum_{i=1}^{x} \sum_{j \neq i}^{x} \psi(h_i, h_j) +$$

$$\frac{1}{y(y-1)} \sum_{i=1}^{y} \sum_{j \neq i}^{y} \psi(\tilde{h}_i, \tilde{h}_j) - \frac{2}{xy} \sum_{i=1}^{x} \sum_{j=1}^{y} \psi(h_i, \tilde{h}_j),$$

which is a sum of two U-statistics and a sample average. Following (Anderson, Hall, and Titterington 1994), we would like to use asymptotic distribution of MMD_u^2 under null hypothesis for the hypothesis testing, by assuming that P and P_{DFD} are identical. Hence if we generate two new data samples from the aggregated data samples after random shuffle, the MMD_u^2 should not change. We can construct null distribution by re-shuffling the aggregated data samples and recomputing the MMD_u^2 a lot of times. Given a significance level α, if MMD_u^2 is so large as to be outside the $1-\alpha$ quantile of the null distribution, we can reject the null hypothesis, otherwise we accept it.

5 Experiments

In this section, we design synthetic and real-world experiments to validate our methodology against the following questions:

QS1 When existing latent structural heterogeneity, will the classical network representation model like node2vec perform fairly across different subgroups?

QS2 Can our method effectively measure fairness of network representation considering structural heterogeneity in subgroups?

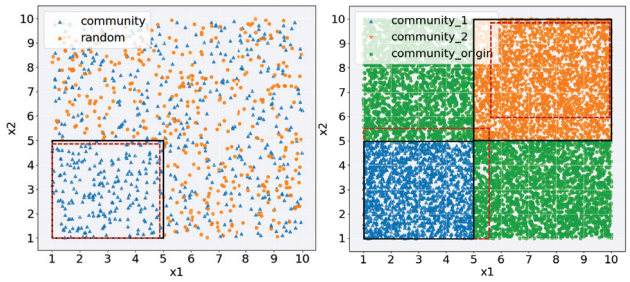
QS3 Are the fairness measurement scores reported by our method significant enough comparing with the random noises?

The most difficult problem for evaluating our methods is the missing of ground truth. For an observational dataset, we do not know whether there is structural heterogeneity and consequently we cannot know whether we can correctly measure the fairness. To overcome this, we design experiments with regard to synthetic data generated by controlling the dependencies between descriptive variables and the network structures. By doing this, the experiments can evaluate the performance of our method by comparing with the ground truth. For real-world datasets, we will never know the ground truth, but the statistical test can help us to evaluate the methods against the random baselines. Qualitative and visual analysis can be used to show the effectiveness of the discoveries.

5.1 Datasets

Synthetic datasets with ground truth As synthetic datasets, we employ modified versions of the two datasets from (Girvan and Newman 2002). The two datasets are called *Karate* and *Football*. We keep the original nodes and community label and drop all the connections. The generating process of the synthetic datasets is governed by following parameters: the number of records N, the number of descriptive variables K, the set of nodes V, and the set of ground truth labels Y indicating communities. We propose a randomized technique to model the dependencies between target variables v_d, v_o and descriptive variables $x_1, \ldots, x_k$. Two kinds of heterogeneous structures are generated: one is community structure in subgroups against uniform distribution of edges in global, another is core-periphery structure (Borgatti and Everett 2000). We visualize two examples 'KarateX4n10k' (K=4, N=$10,000$, $|V|$=34) and 'FootballX4n10k' (K=4, N=$10,000$, $|V|$=115) in Figure 2. In Figure 2a, triangles represent the edges inside communities and dots represent uniform sampled edges between any pair of nodes. We can see that blue triangles distribute uniformly except in the black rectangle. In the ground truth subgroup, the edges only exist in the local community. In Figure 2b, we synthesize a simple core-periphery structure. This is one of the simplest global role structures which consists of dense and cohesive core nodes as well as sparse and unconnected periphery nodes.

Real-world datasets As real-world datasets, two kinds of data are used for the experiments: (1) the original edge connections; and (2) extra data about the contextual information. We collect the original edge connections including 'New York Taxi' (http://www.nyc.gov/html/tlc/) (K=33, N=$1,013,845$, $|V|$=265) and 'Sharing Bike' (https://datasf.org/opendata/) (K=27, N=$983,000$, $|V|$=70), as well as the contextual information, e.g. weather records (https://www.ncdc.noaa.gov/) and taxi information. By choosing these



(a) KarateX4n1k. (b) FootballX4n10k.

Figure 2: Randomized synthetic datasets with ground truth. Rectangles with solid lines denote ground truth subgroups. Rectangles with dash lines denote the subgroups reported by our method.

two datasets, we would like to show how network representation model can be biased by attributes like weather conditions. Consequently the downstream tasks (e.g. transportation prediction on specific weather conditions) could also be biased using the representations learned from the whole data. From these experiments we show that study for fairness of network representation has broad application fields.

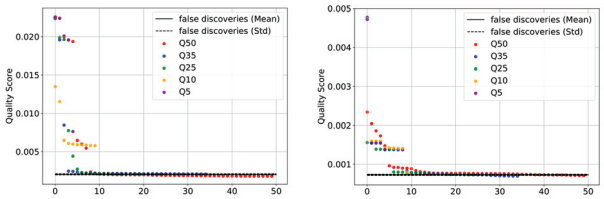
5.2 Implementation details

For the implementation of node representation learning, we build the algorithm based on Node2vec (Grover and Leskovec 2016). For each candidate subgroup, we construct the graph with edges covered by that subgroup and use random walk algorithm considering the aggregated edge weights to generate the training labels. After getting the node representations, we compare them with node representations learned from the whole data. To explore the attribute space with exponential amounts of subgroups, we use beam search guiding by the quality score heuristically. The beam search algorithm is built based on (Duivesteijn, Feelders, and Knobbe 2016, Algorithm 1). We set the beam width to 5 and depth to 2. All the experiments are conducted on Linux computing clusters with CPU: 2x Intel Xeon @ 2.1GHz and RAM: 1024GB.

5.3 Experiments on Synthetic Data

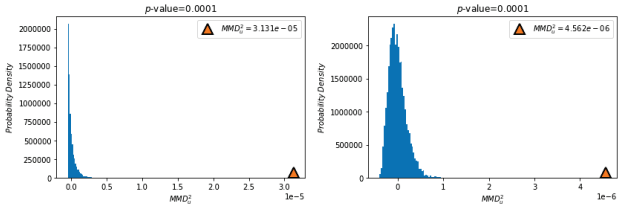
To validate our method against QS1 and QS2, we conduct experiments on the two synthetic datasets with different settings mainly by varying parameter Q, which indicates how many subgroups we are going to report. The top-5 subgroups are reported in Table 2. As shown in Figure 2, our algorithm can discover the pre-imposed structures with good accuracy.

The subgroups we found cannot always be precisely the ground truth. The rectangles with black solid lines and the rectangles with red dot lines are slightly mismatching (cf. Figure 2). There might be two reasons for that. On the one hand, we employ a 8-bin equal-width binning strategy to partition the space of descriptive variables denoted by continuous numerical values. On the other hand, we prune the result set based on overlapping coverage to reduce redundant discoveries. Hence, we plan to evaluate more about the



(a) KarateX4n10k. (b) FootballX4n10k.

Figure 3: Comparisons of quality score distributions.



(a) KarateX4n10k. Q=10. (b) FootballX4n10k. Q=10.

Figure 4: Visualization of null distribution and MMD_u^2 on KarateX4n10k and FootballX4n10k datasets.

predictive ability of our method. According to the known label of each edge, we can calculate averaged number of edges covered by discovered subgroups to build the confusion matrix. We choose true positive rate (TPR) and positive predictive value (PPV) as the evaluation indicators.

Table 3 displays the results; larger TPR and PPV indicate better results. We can see that for the same dataset, with the increasing of Q, MMD_u^2, TPR and PPV decrease. One reason for this phenomenon is that the forced diversity of discovered subgroups works against identification of the single ground truth subgroup. Another reason is that larger Q allows for subgroups with lower qualities, so that some records without label of ground truth are discovered by our method. We also notice that the PPV of finding subgroups by our method are always larger than 50%, which shows that our method can reliably retrieve ground-truth subgroups.

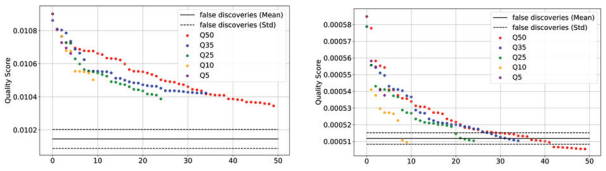
In order to validate our method against QS3, we run our algorithm on the randomly shuffled datasets for 100 times to generate negative samples. In Figure 3, we plot the quality scores in different experiments with Q ranging from 5 to 50, as well as the quality scores from negative samples. We can see that there is a large gap between quality scores of reported subgroups and the false discoveries. One reason is that with synthetic algorithm, we impose very different structural properties. Also we noticed that there are many low ranked subgroups dropping into the region of false discoveries. The reason is that the number of pre-imposed discriminated subgroups are less than the Q. Then we conduct the hypothesis testing to investigate whether the differences between our discoveries and the false discoveries are significant enough. In Figure 4, we visualize the null distribution and report p-value with $Q = 10$ on KarateX4n10k and Foot-

KarateX4n10k			FootballX4n10k		
D	$\varphi_{\text{MLSD}}(D)$	$\frac{\lvert D\rvert}{N}$	D	$\varphi_{\text{MLSD}}(D)$	$\frac{\lvert D\rvert}{N}$
$x_1 \leq 4.86 \wedge x_2 \leq 4.86$	.0225	.188	$x_2 \geq 6.14 \wedge x_1 \geq 4.86$	.0047	0.244
$x_1 \leq 3.57 \wedge x_2 \leq 4.86$	.0224	.188	$x_2 \geq 6.14 \wedge x_1 \geq 6.14$	.0015	0.182
$x_1 \leq 4.86 \wedge x_2 \leq 3.57$	.0201	.128	$x_2 \leq 4.86 \wedge x_1 \leq 4.86$	.0015	0.182
$x_1 \leq 3.57 \wedge x_2 \leq 3.57$	.0196	.123	$x_1 \leq 4.86 \wedge x_2 \leq 3.57$	.0014	0.124
$x_1 \leq 6.14 \wedge x_2 \leq 4.86$	.0076	.249	$x_1 \leq 3.57 \wedge x_2 \leq 4.86$	.0014	0.124

Table 2: Top-5 subgroups discovered on KarateX4n10k. The higher $\varphi_{\text{MLSD}}(D)$, the more unfair. $\frac{\lvert D\rvert}{N}$ indicates the coverage of subgroups.

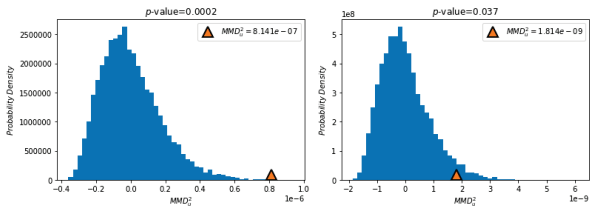
KarateX4n10k			FootballX4n10k		
Q	TPR	PPV	Q	TPR	PPV
5	.61	.94	5	.69	1.0
10	.40	.86	10	.53	.96
25	.36	.71	25	.44	.52
35	.36	.65	35	.28	.51
50	.33	.50	50	.28	.50

Table 3: Experimental results on synthetic datasets. The higher TRP and PPV the better.



(a) Quality distribution bike. (b) Quality distribution taxi.

Figure 5: Quality score comparisons on dataset Sharing Bike and New York Taxi.



(a) Bike. Q=10. (b) Taxi. Q=10.

Figure 6: Visualization of null distribution and MMD_u^2 on bike and taxi datasets.

ballX4n10k. As we can see intuitively, the MMD_u^2 is far from null distribution. We can be confident that our method can beat false discoveries generated from random baselines. We also noticed that based on the p-values we can reject the null hypothesis at 1% significance level.

5.4 Experiments on Real-world Datasets

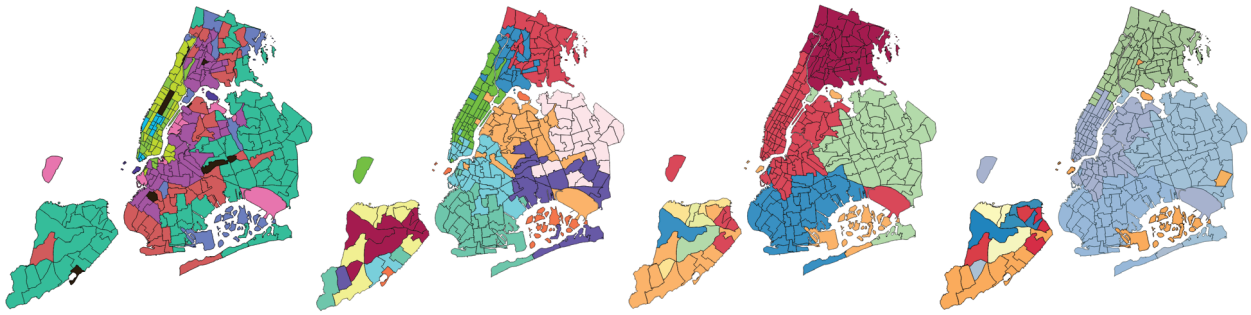
Similar experiments are conducted on the real-world datasets, except calculating TRP and PPV due to the reason that we do not know the ground truth. In Figure 5, we plot the quality scores of discovered subgroups in different experimental settings with Q ranging from 5 to 50. We can see that in the real-world datasets, the quality decreases smoothly than the synthetic. One reason might be that in the real-world datasets, there are many kinds of combinations between structural properties and descriptive variables. Another reason might be that the attribute space and number of edges are much larger than the synthetic datasets so that the performance of network representation models are more diverse. As we can see in Figure 6, the MMD_u^2 and p-values give us confidence to believe that there are significant differences between the subgroups reported by our method and the false discoveries. In Table 4, we report the top-5 subgroups in both datasets. We can see from the descriptions that the weather conditions and urban regions are highly related with the heterogeneous structures. This indicates that the decision models might be more vulnerable and discriminated under such conditions.

Empirical Clustering Analysis To further explore these results, we conduct clustering on taxi zones in New York using k-means algorithm with the learned representations from taxi transitions. We use the discovered subgroups above and the whole dataset as the input to train representations for each taxi zone. On the one hand, we would like to see how these clusters are different between reported subgroups and the whole dataset. On the other hand, we would like to see how the representations of taxi zones are changing with the changing of descriptive variables.

To conduct this comparison, we employ the land use data in New York (https://zola.planning.nyc.gov/) as a reference of the ground truth. The assumption is that taxi zones with similar land use types are similar to each other. Based on this assumption, we count the land use types in each taxi zone, and compute the distribution of land use types as the representation of each taxi zone. We visualize these clustering results in Figure 7. By comparing those clusters in Figure 7a with the clusters learned on the whole dataset (cf. Figure 7b), we found the similarities between taxi zones can be preserved relatively well. In Figure 7c, John F. Kennedy International Airport shows different role with nearby zones, while it shows the same role with the Manhattan area. In Figure 7d, we can see that for 'passenger $>$ 5', many zones that are distinguished in previous subgroups become more

Dataset	D	$\varphi_{\text{MLSD}}(D)$	$\frac{\lvert D\rvert}{N}$
Sharing Bike	MaxHumidity <= 74.0∧ZipCode ! = '10010'	.01090	.194
	MinTemperatureF > 50.0∧MaxTemperatureF > 70.0'	.01081	.232
	MaxHumidity <= 74.0∧ZipCode ! = '7050'	.01073	.194
	MaxHumidity <= 74.0∧ZipCode ! = '77450'	.01069	.194
	MaxHumidity <= 74.0∧ZipCode ! = '19119'	.01066	.194
New York Taxi	month > 7.0∧PaymentType <= 1.0	5.85e-4	.211
	TMIN > 61.0∧PickupHour <= 14:00	5.58e-4	.126
	month > 7.0∧AWND <=5.24	5.54e-4	.272
	month > 7.0∧TMIN > 42.0	5.41e-4	.279
	month > 7.0∧TMAX > 54.0	5.38e-4	.300

Table 4: Experiments on real-world datasets. Higher $\varphi_{\text{MLSD}}(D)$ means more unfair.



(a) Land Use. (b) Ω. (c) D_1. (d) Passenger> 5.

Figure 7: Taxi zone clusters with representations. D_1:'SNWD$>$0.0∧AWND$\leq$ 7.86'.

similar. These results empirically show the structural heterogeneity in different subgroups. For fair decision making, a network representation model should tackle this heterogeneity to learn fair as well as informative representations.

6 Conclusions

In this paper, we study an important problem: fairness in network representation by latent structural heterogeneity in observational data. We argue that the structural heterogeneity in networks can bias the network representation models across subgroups, which will prevent us from building fair decision making models for downstream tasks like node classification or link prediction. However, the unknown distribution of structural heterogeneity raises new challenges for fairness measurement. Pre-defined groups with sensitive variables are not proper for overcoming the new challenges, and statistical parity with regard to decision variable cannot be helpful for comparing the multi-degree interactions between node representations. We analyze the connections between the structural properties and the node representations in networks. Then we design a framework to compare the node representations learned from subgroups with the node representations learned from the whole data. The differences between them indicate that the structural properties in subgroups are ignored by the network representation model. The higher the difference, the more unfair the model is on those subgroups. The discovery process is automatically guided by a search algorithm defined over the description space, with a quality measure over the learned node representations, called Mean Latent Similarity Discrepancy (MLSD). We evaluate the statistical significance of the discovered subgroups by applying a kernel two-sample test. To validate the effectiveness of our method, we use randomization techniques to generate synthetic datasets with ground truth. This allows us to evaluate the performance of our method quantitatively and qualitatively. In future work, we will integrate the representation learning and subgroup discovery process to generate fair and informative node representations for downstream decision making applications.

References

Anderson, N. H.; Hall, P.; and Titterington, D. M. 1994. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis* 50(1).

Atzmueller, M. 2015. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5(1).

Batagelj, V., and Brandes, U. 2005. Efficient generation of large random networks. *Physical Review E* 71(3).

Bendimerad, A. A.; Plantevit, M.; and Robardet, C. 2016. Unsupervised exceptional attributed sub-graph mining in urban data. In *Proc. ICDM*.

Borgatti, S. P., and Everett, M. G. 2000. Models of core/periphery structures. *Social networks* 21(4).

Brennan, T.; Dieterich, W.; and Ehret, B. 2009. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior* 36(1):21–40.

Chen, I.; Johansson, F. D.; and Sontag, D. 2018. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002*.

Duivesteijn, W., and Knobbe, A. 2011. Exploiting false discoveries–statistical validation of patterns and quality measures in subgroup discovery. In *Proc. ICDM*.

Duivesteijn, W.; Knobbe, A.; Feelders, A.; and van Leeuwen, M. 2010. Subgroup discovery meets Bayesian networks–an exceptional model mining approach. In *Proc. ICDM*. IEEE.

Duivesteijn, W.; Feelders, A. J.; and Knobbe, A. 2016. Exceptional model mining. *Data Mining and Knowledge Discovery* 30(1).

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proc. ITCS*, 214–226. ACM.

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

Girvan, M., and Newman, M. E. 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences* 99(12).

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13(Mar).

Grover, A., and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proc. KDD*.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323.

Herrera, F.; Carmona, C. J.; González, P.; and del Jesús, M. J. 2011. An overview on subgroup discovery: foundations and applications. *Knowl. Inf. Syst.* 29(3):495–525.

Jin, R.; Lee, V. E.; and Hong, H. 2011. Axiomatic ranking of network role similarity. In *Proc. KDD*.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*.

Khandani, A. E.; Kim, A. J.; and Lo, A. W. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34(11):2767–2787.

Klösgen, W. 1996. Explora: A multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*. 249–271.

Korolyuk, V. S., and Borovskich, Y. V. 2013. *Theory of U-statistics*, volume 273. Springer Science & Business Media.

Lemmerich, F.; Becker, M.; Singer, P.; Helic, D.; Hotho, A.; and Strohmaier, M. 2016. Mining subgroups with exceptional transition behavior. In *Proc. KDD*.

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

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2019. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proc. FAT**, 349–358. ACM.

Mahoney, J. F., and Mohen, J. M. 2007. Method and system for loan origination and underwriting. US Patent 7,287,008.

Mampaey, M.; Nijssen, S.; Feelders, A.; Konijn, R.; and Knobbe, A. 2015. Efficient algorithms for finding optimal binary features in numeric and nominal labeled data. *Knowledge and Information Systems* 42(2):465–492.

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

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*.

Müller, A. 1997. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability* 29(2).

Pearl, J. 2017. Detecting latent heterogeneity. *Sociological Methods & Research* 46(3).

Pei, Y.; Zhang, J.; Fletcher, G.; and Pechenizkiy, M. 2018. Dynmf: role analytics in dynamic social networks. In *Proceedings of the 27th IJCAI*.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk: Online learning of social representations. In *Proc. KDD*.

Ribeiro, L. F.; Saverese, P. H.; and Figueiredo, D. R. 2017. struc2vec: Learning node representations from structural identity. In *Proc. KDD*.

Song, J.; Kalluri, P.; Grover, A.; Zhao, S.; and Ermon, S. 2018. Learning controllable fair representations. *arXiv preprint arXiv:1812.04218*.

Todorovski, L.; Flach, P.; and Lavrač, N. 2000. Predictive performance of weighted relative accuracy. In *Proc. ECMLPKDD*, 255–264. Springer.

Wrobel, S. 1997. An algorithm for multi-relational discovery of subgroups. In *Proc. PKDD*, 78–87.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2015. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*.

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

Zhao, H., and Gordon, G. J. 2019. Inherent tradeoffs in learning fair representation. *arXiv preprint arXiv:1906.08386*.