Fairness-Aware Graph Sampling for Network Analysis

Farzan Masrour, Francisco Santos, Pang-Ning Tan, Abdol-Hossein Esfahanian
Department of Computer Science and Engineering, Michigan State University
Emails:{masrours, santosf3, ptan, esfahanian}@msu.edu

***Abstract*—Network sampling is the task of selecting a subset of nodes and links from a network in a way that preserves its topological properties and other user requirements. This paper investigates the problem of generating an unbiased network sample that contains balanced proportion of nodes from different groups. Creating such a representative sample would require handling the trade-off between ensuring structural preservability and group representativity of the selected nodes. We present a novel max-min subgraph fairness measure that can be used as a unifying framework to combine both criteria. A greedy algorithm is then proposed to generate a fair and representative sample from an initial set of target nodes. A theoretical approximation guarantee for the output of the proposed greedy algorithm based on submodularity and curvature ratios is also presented. Experimental results on real-world datasets show that the proposed method will generate more fair and representative samples compared to other existing network sampling methods.**

***Index Terms*—Sampling; network; fairness**

I. Introduction

Networks are powerful representation for modeling interactions between entities in a complex social system. Given the massive size of online social networks, performing even simple analysis can be expensive. Besides the high computational costs, many online social networks are only accessible through Web crawling or the use of APIs. Such accesses are often throttled by restrictions on the number of queries or rate limits imposed by the data providers. Accessing the whole network becomes near impossible for researchers, which makes sampling an essential task for collecting network data.

An obvious sampling goal is to ensure that the sample preserves the topological properties of the entire network [1]–[3]. This objective will hereforth be referred to as *structural preservability* of the sample. In graph sampling literature [4], [5], numerous network topological measures have been used to characterize structural preservability. These properties can be generally categorized into two groups: (1) vector properties such as the distribution of node degrees, clustering coefficients, and eigenvalues, and (2) scalar properties such as average degree, network diameter, and average clustering coefficient. The former can be assessed using probabilistic distance measures such as Kullback–Leibler divergence and Kolmogorov-Smirnov statistic whereas the latter can be evaluated by computing the normalized root mean square error (NRMSE) [6] between the scalar properties obtained from the sample and those obtained from the entire network.

Each node in a social network can be characterized by its attributes. Some node attributes may define certain groups (e.g., gender, race, or age group) of interest to network researchers. These are known as the protected attributes. Thus, an alternative sampling goal would be to preserve the distribution of such protected attribute values in the sample [6]. This objective will be referred to herein as *group representativity*. Given its broad range of applications, the importance of fair network sampling cannot be overly emphasized. If the sampled network is biased, this will adversely affect the results of downstream mining tasks. For example, Wagner et al. [7] showed that uninformed sampling may lead to biased estimation of node centrality values and unfair ranking of nodes from minority groups in a social network. The authors noted that an ideal sample should not "systematically rank nodes of one group higher and nodes of the other group lower than expected." Nevertheless, they did not present a fair sampling method that will overcome the limitations of existing algorithms.

The main challenge in fair network sampling is to combine the structural preservability and group representativity objectives in a principled way and to design an algorithm that optimizes for both. In this paper, we develop an approach that measures structural preservability by comparing the centrality measures of the nodes in the sampled network to their values in the original network. Fairness of the sampled network, which corresponds to the group representativity objective, is given by a max-min subgroup fairness criterion [8], defined in terms of the worst-case structural preservability value among all the subgroups of the protected attribute. A greedy algorithm is then proposed to obtain an approximate solution for the max-min subgroup fairness criterion. A systematic evaluation of the proposed algorithm on several real-world data demonstrates its effectiveness relative to other network sampling methods.

II. Preliminaries

Let $G =< V, E, X >$ be an attributed network, where V is the set of nodes, $E \subseteq V \times V$ is the set of links, and X is the set of node attributes. We assume X can be partitioned into protected attributes, $X^{(p)}$, and unprotected attributes, $X^{(u)}$. Let $\{P_1, \ldots, P_K\}$ be the partitions of the nodes in V based on the values of the protected attributes in $X^{(p)}$, where K is the number of distinct combination of its values.

Consider a pair of networks, $G_1 =< V_1, E_1, X_1 >$ and $G_2 =< V_2, E_2, X_2 >$. We say that G_1 is a subgraph of G_2, denoted as $G_1 \subseteq G_2$, if $V_1 \subseteq V_2$ and $E_1 \subseteq E_2$. If E_1 includes all the links in G_2 that have endpoints in V_1, then G_1 is an induced subgraph of G_2.

Definition 1 (Network Sampling)**.** *Given an attributed network $G =< V, E, X >$ and a sample size n, the network sampling task is to extract an induced subgraph $G_s^* =< V_s^*, E_s^*, X_s^* >$ of G such that:*

$$G_s^* = argmax_{G_s \subset G, |V_s|=n} \mathbf{o}(G_s) \tag{1}$$

where $\mathbf{o}$ is an objective function that considers both structural preservability and group representativity of the sample.

A. Structural Preservability Objective

A natural starting point for measuring the ability of a sample to preserve the topological properties of a network is at the individual node level. More specifically, node centrality measures [9]–[11] such as degree, closeness, clustering coefficient, Katz index, and PageRank have been widely used to gauge the importance of a node in a network. By comparing the centrality measure of each node computed from the sample to its corresponding value from the original network, the structural preservability of a sample can therefore be evaluated.

Definition 2 (Centrality Ratio)**.** *Given a network $G =< V, E, X >$ and a subgraph G_s, the centrality ratio μ of a node v in G_s is defined as*

$$\mu(v, G_s) = \frac{C(v, G_s)}{C(v, G)} \tag{2}$$

where C corresponds to a node centrality measure.

Intuitively, if μ is close to 1, then the sample G_s preserves the centrality measure of node v in the original network G. The structural preservability objective aims to find a subgraph G_s of order n that minimizes the following function:

$$\mathbf{o}_{struct}(G_s) = \sum_{v \in G} |\mu(v, G_s) - 1| \tag{3}$$

One potential caveat of using Equations (2) and (3) is that they both require knowledge of the entire network in order to compute the denominator term $C(v, G)$. If the sampling algorithm is restricted to have access only to the sampled network (e.g., while crawling the network) without any prior knowledge about the entire network, it would not be possible to compute $\mu(v, G_s)$. To circumvent this issue, we consider the following relaxed centrality ratio measure instead:

$$\tilde{\mu}(v, G_s) \equiv C(v, G_s) \tag{4}$$

For greedy algorithms that incrementally expand the sampled subgraph by adding one node at a time, the relaxed measure provides a good approximation to the structural preservability objective in (3) when knowledge about the entire network is unavailable. The true centrality ratio given by Equation (2) can be used when evaluating the performance of such algorithms.

For a greedy sampling algorithm, the following are 3 desirable properties of the centrality and centrality ratio measures:

- P1: $\tilde{\mu}(v, G_s) = 0$ if $v \notin V_s$.
- P2: $\forall v \in V_s : \tilde{\mu}(v, G_s) < \infty$ even if G_s is not a connected graph.
- P3: $\forall v \in V_1 : \tilde{\mu}(v, G_1) \leq \tilde{\mu}(v, G_2)$ if $G_1 \subseteq G_2$.

The first property ensures that the centrality ratio can be computed using information from the sample G_s alone. The second property ensures that the centrality ratio remains bounded even when the subgraph has multiple connected components. The third property, which is analogous to the score monotonicity axiom for node centrality defined in [10], implies that the centrality ratio of the subgraph should be monotonically non-decreasing as the order ($|V_s|$) of the subgraph increases.

Our algorithm considers harmonic centrality [10] as the node centrality measure. The harmonic centrality of a node u with respect to graph G, $H(u, G)$, is defined as:

$$H(u, G) \equiv \sum_{x \in V \setminus \{u\}} H(u, x) \; = \sum_{x \in V \setminus \{u\}} \frac{1}{d_G(u, x)}, \tag{5}$$

where $d_G(u, x)$ denotes the shortest path distance between u and x in G. We choose harmonic centrality for several reasons. First, unlike other path-based node centrality measures such as closeness and betweenness, which are restricted to connected graphs, harmonic distance is applicable to both connected and disconnected graphs (property P2). Furthermore, the measure is intuitive as it considers the relative influence of the nodes in a network by giving higher weights to nodes that are closer than those located further away. Another advantage of using harmonic centrality is that it is only one of two popular measures (besides PageRank) that is strictly rank monotone [11], i.e., adding a new edge to a node will not demote its rank relative to other lower-ranked nodes in the network.

Lemma 1. *Harmonic centrality satisfies the desired properties P1, P2, and P3 of a greedy, graph sampling algorithm.*

Proof: For property P1, it is easy to show that $\tilde{\mu}(v, G_s) = 0$ for $v \notin V_s$ since $\forall u \in V_s : d_G(v, u) = \infty \Rightarrow H(v, u) = 0$. For property P2, if there is no path between nodes u and x, then $d_G(u, x) = \infty$. Thus, $H(u, x) = 1/d_G(u, x) = 0$ when u and x belong to different connected components. The corresponding term in the sum given in Equation (5) will be zero, which means $H(u, G)$ is bounded even if G is not a connected graph. To prove the third property, let $G_1 =< V_1, E_1 >$ and $G_2 =< V_2, E_2 >$. If $G_1 \subseteq G_2$ then $\forall v, u \in V_1 : d_{G_1}(v, u) \geq d_{G_2}(v, u)$. As a result,

$$\forall v, u \in V_1 : \frac{1}{d_{G_1}(v, u)} \leq \frac{1}{d_{G_2}(v, u)}.$$

Furthermore, since $V_1 \subseteq V_2$, we have $\forall v \in V_1$:

$$\begin{aligned} H(v, G_2) \;&=\; \sum_{u \in V_2 \setminus \{v\}} \frac{1}{d_{G_2}(v, u)} \geq \sum_{u \in V_1 \setminus \{v\}} \frac{1}{d_{G_2}(v, u)} \\ &\geq \; \sum_{u \in V_1 \setminus \{v\}} \frac{1}{d_{G_1}(v, u)} = H(v, G_1) \end{aligned}$$

Thus, the third property holds for Equation (4). □

B. Group Representativity Objective

The group representativity objective ensures that the sample contains adequate representation from different groups of the protected attribute(s). To address the challenge of combining

structural preservability with group representativity, we introduce a ***max-min subgraph fairness*** criterion, which is inspired by the minimax Pareto fairness concept proposed in [8] for satisfying group fairness. Specifically, our max-min subgraph fairness criterion evaluates the average centrality ratio of each group and uses the minimum value as its fairness score.

Let $\mathbb{P} = \{P_1, P_2, \cdots, P_K\}$ be a partitioning of the nodes based on the protected attribute(s), $X^{(p)}$. For each group $P_i \in \mathbb{P}$, we define the following group centrality ratio measure:

Definition 3 (Group Centrality Ratio)**.** *Given a network $G =< V, E, X >$ and centrality ratio μ, the group centrality ratio for the node group $P_i \in \mathbb{P}$ is*

$$\sigma_i(G_s) = \frac{1}{|P_i|} \sum_{u \in P_i} \mu(u, G_s) \tag{6}$$

where $|P_i|$ denotes cardinality of the set of nodes in P_i.

For sampling algorithms without full access to the entire network, we replace μ by $\tilde{\mu}$ in the above definition. Furthermore, since harmonic centrality satisfies property P1, the group centrality ratio can be computed efficiently as follows:

$$\sigma_i(G_s) = \frac{1}{|P_i|} \sum_{u \in P_i \cap V_s} \mu(u, G_s) \tag{7}$$

Definition 4 (Subgraph Fairness Criterion)**.** *Given a network sample $G_s =< V_s, E_s, X_s >$, a partitioning of node groups, $\mathbb{P} = \{P_1, P_2, \cdots, P_K\}$, and a group centrality ratio function σ, we define the fairness function for a subgraph G_s as:*

$$\mathbf{o}_{fair}(G_s) \equiv \sigma^{min}(V_s) = \min_{1 \le i \le K} \{\sigma_i(G_s)\} \tag{8}$$

Our network sampling goal is to maximize the fairness criterion defined in Equation (8). Replacing the measure into the objective function in Equation (1), our max-min subgraph fairness sampling objective is to find a subgraph G_s^* such that:

$$G_s^* = \text{argmax}_{G_s \in G, |V_s| = n} \min_i(\sigma_i(G_s)) \tag{9}$$

Lemma 2. *If $\tilde{\mu}$ satisfies the property P3 stated in Lemma 1, then $\sigma^{min}(V_s)$ is a monotonically non-decreasing function of the order of the subgraph.*

Proof: Let $\sigma^{min}(V_s) = \min_{1 \le i \le K}(\sigma_i(G_s))$, where K is the number of group partitions. In order to show σ^{min} is a monotonically non-decreasing function, we have to show that $\sigma^{min}(V_1) \le \sigma^{min}(V_2)$ if $G_1 \subseteq G_2$, where $G_i =< V_i, E_i >$. Since $\tilde{\mu}$ satisfies the property P3 in Lemma 1 and σ_i is a summation over μ, therefore σ_i must be a monotonically nondecreasing function for all $1 \le i \le K$. Furthermore, as σ^{min} is the minimum value of σ_i over i, it must also be monotonically nondecreasing, which completes the proof. □

III. GFNS: Greedy Fair Network Sampling

Our proposed greedy algorithm to address the max-min subgraph fairness criterion given in (9) is based on the following notion of marginal gain of a set of nodes $A \subset V$.

Definition 5 (Marginal Gain)**.** *Given a network $G = (V, E)$, a subgraph $G_s = (V_s, E_s)$, and a set of nodes $A \subset V \setminus V_s$, the marginal gain $\delta(.)$ of adding A to V_s is*

$$\delta(A|G_s) = \sigma^{min}(V_s \cup A) - \sigma^{min}(V_s) \tag{10}$$

where $\sigma^{min}(V_s) = \min_{1 \le i \le K}\{\sigma_i(G_s)\}$ as defined in (8)*.*

A greedy algorithm can be developed to optimize (9) by incrementally adding a node v and its corresponding edges in E into the sample G_s in a way that maximizes the marginal gain. However, computing the harmonic centrality can be very expensive when the sampled graph is large. To improve its efficiency, we present the following fast implementation of our greedy algorithm based on a reference set of target nodes.

Definition 6 (Target Set)**.** *Given a network $G =< V, E, X >$ and a partitioning of the node groups $\mathbb{P} = \{P_1, P_2, \cdots, P_K\}$, where $V = \cup_{i=1}^{K} P_i$, the target set $T = \{T_1, \ldots, T_k\}$ is a set of node subsets such that $T_i \subset P_i$, for $1 \le i \le K$.*

We will use the target set to compute the following approximate group centrality ratio:

$$\tilde{\sigma}_i(G_s) = \frac{1}{|T_i|} \sum_{u \in T_i} \mu(u, G_s) \tag{11}$$

for each group. The marginal gain in Equation (10) can then be computed using the approximate group centrality ratio instead. Our greedy algorithm is summarized below.

Greedy Fair Network Sampling (GFNS) Algorithm

Input: network G, sample size n, and target set T.
Output: sampled subgraph, $G^{(n)}$.

$G^{(|T|)} \leftarrow$ Induced-subgraph(T, G).
for t = $|T| + 1$ to $n - 1$ **do**
  $\chi \leftarrow \{u \mid (u, v) \in G, u \in V \setminus V^{(t-1)}, v \in V^{(t-1)}\}$
  $v^* \leftarrow \text{argmax}_{v \in \chi} \delta(v|G^{(t-1)})$
  $G^{(t)} \leftarrow$ Induced-subgraph($V_{t-1} \cup \{v^*\}, G$)
end for

A. Theoretical Bounds on Greedy Approximation

Note that the max-min subgraph fairness criterion is not a submodular function. However, as shown in Lemma 2, it is a monotonically nondecreasing function. This allows us to use the result of [12] to obtain a theoretical bound on the greedy approximated solution. Before presenting the main theorem, we first introduce some notations and definitions. Let $G^{(0)}, G^{(1)}, \cdots, G^{(n)}$ be the sequence of subgraph samples generated from a network G, where $G^{(t)} =< V^{(t)}, E^{(t)} >$ and $|V^{(t)}| = t$.

Definition 7 (Submodularity ratio [13])**.** *The submodularity ratio of a function σ is the largest scalar γ such that*

$$\sum_{u \in A \setminus V^{(t)}} \delta(u|G^{(t)}) \ge \gamma \delta(A|G^{(t)}), \quad \forall A \subset V : |A| = n$$

where $\delta(\cdot)$ is the marginal gain defined for σ and $t \in \{0, 1, \ldots n-1\}$.

TABLE I: Statistics of network data used for experiments.

Network	#nodes	#edges	CC	protected feature
Facebook [14]	4,039	88234	0.6055	gender
Tagged [15]	71,127	71,265	0.0005	gender
German [16]	1,000	24,970	0.3801	gender
Credit [17]	30,000	2,174,014	0.6466	Age

Note that, for a non-decreasing function σ, $\gamma \in [0, 1]$ [12]. Furthermore, the function is submodular if and only if $\gamma = 1$.

Definition 8 (Greedy curvature [12])**.** *The greedy curvature of a function σ is the smallest scalar α such that*

$$\delta(v_t|G^{(t-1)} \cup A) \geq (1-\alpha)\delta(v_t|G^{(t)}),\ \forall A \subset V : |A| = n$$

where $\delta(\cdot)$ is the marginal gain defined for σ and $t \in \{0, 1, \ldots n-1\}$.

Theorem 1. *Let σ be the group centrality measure defined in Equation* (6) *and $\delta(\cdot)$ be its marginal gain with submodularity ratio and greedy curvature as defined in Definitions 7 and 8, respectively. The proposed greedy fair network sampling algorithm has the following approximation guarantee*

$$\begin{aligned} \sigma^{\min}(V^{(n)}) &\geq \frac{1}{\alpha}\Big[1 - \Big(\frac{n-\alpha\gamma}{n}\Big)^n\Big]\sigma^{\min}(V^*) \\ &\geq \frac{1}{\alpha}(1 - e^{-\alpha\gamma})\sigma^{\min}(V^*) \end{aligned}$$

where $G^{(n)} =< V^{(n)}, E^{(n)} >$ is the output of the greedy algorithm and $G^* =< V^*, E^* >$ is the optimum solution.

The proof of the theorem can be shown using Lemma 2 of this paper and by applying Theorem 1 of [12].

IV. Experimental Evaluation

This section describes the experiments performed to evaluate the efficacy of our proposed sampling algorithm. All the code and datasets for our experiments are available at https://github.com/frsantosp/FairSampling.

A. Experimental Setup

We performed experiments on 4 real-world datasets, whose properties are summarized in Table I. Gender is chosen as the protected attribute for the first 3 datasets while age is the protected attribute for the credit default dataset. We evaluated the structural preservability of the sampling algorithm using the following metrics:

- **Degree distribution distance (Ddist):** Following the approach in [1], we compare the degree distribution of the sampled graph against the original network using the Kolmogorov-Smirnov (K-S) statistic:

  $$Ddist(G_s) = \sup_d |F_S(d) - F(d)|$$

  where F_S and F are the cumulative distribution function (CDF) for the degree distributions of the sampled graph and the original networks, respectively.
- **Clustering Coefficient(δ-CC):** We compare the difference in average clustering coefficients of the nodes in the original network to the sampled graph as follows:

  $$\delta\text{-CC} = \Big|\frac{1}{|V|}\sum_{v\in V}\frac{\lambda_G(v)}{\tau_G(v)} - \frac{1}{|V_s|}\sum_{v\in V_s}\frac{\lambda_{G_s}(v)}{\tau_{G_s}(v)}\Big|$$

  where $\lambda_G(v)$ is the number of triangles in G involving the node v while $\tau_G(v)$ is the corresponding number of open triangles (i.e., subgraphs with 2 links and 3 nodes) in G with v being the bridge between two other nodes.
- **Harmonic:** the average centrality ratio (see Equation (2)) of all the nodes in the sampled network using harmonic distance as centrality measure.

For group representativity, we employed the metrics below:

- **Normalized Cumulative Group Relevance (nCGR)** [7], which measures the extent to which the rank of a node from each group has changed in the sampled graph compared to the original graph. To do this, we first compute the relevance of a node v in a given graph:

  $$rel(v) = \frac{(rank(v))^{-1}}{\sum_{u\in V} rank(v)}$$

  The cumulative protected group relevance [7] for group i is then calculated as follows:

  $$nCGR_i = \frac{\sum_{v\in topk(G_s)\cap P_i} rel(v) + \epsilon}{\sum_{v\in topk(G)\cap P_i} rel(v) + \epsilon}$$

  Note that $nCGR_i$ determines whether the nodes belonging to the group P_i are more or less relevant compared to their expected value in the original network. We set $\epsilon = 0.001$ to avoid division by zero [7] and report the minority group nCGR in our experiments. The closer the nCGR value is to 1, the less biased is the algorithm.
- **min-σ**: The subgraph fairness criterion using the group centrality ratio defined in Equation (8). We consider all the nodes in the sampled subgraph when computing σ.

The following commonly used sampling methods are chosen as baseline algorithms for comparison:

- Random Node Sampling: We consider two variations of this approach: (1) **NS**, which randomly selects a subset of the nodes from a uniform probability distribution and (2) **NSD**, which randomly selects a subset of the nodes with sampling probability proportional to the node degree.
- Breadth/Depth-first search **BFS/ DFS**: Both algorithms start from an initial set of target nodes and iteratively expand the sample based on their graph traversal strategy.
- Random Walk: **RW** starts from a set of seed nodes and expands the sample by simulating a random walk on the network. Fair Random Walk (**FRW**) [18] is a variation of the method that accounts for fairness by partitioning the neighboring nodes into groups based on their protected attribute(s). Each group has equal probability to be chosen as the next node to visit regardless of their cardinality.
- Metropolis-Hastings Random Walk (**MHRW**): MHRW sampling allows the RW to remain at its current node

TABLE II: Performance comparison among various methods in terms of structural preservability (δ-CC, Ddist, and Harmonic) and group representativity (nCGR and min-σ) objectives. Note that FRW and GFNS are fairness-aware sampling methods while the rest are conventional sampling methods. The rank of each method (per evaluation metric) in increasing ($\nearrow$) or decreasing ($\searrow$) order is shown in parenthesis.

	δ-CC ($\nearrow$)	Ddist ($\nearrow$)	Harmonic ($\searrow$)	nCGR ($\nearrow$)	min-σ ($\searrow$)	Average rank
NS	**0.0078+/-0.0000** (1)	0.6159+/-0.0005 (8)	0.3680+/-0.0000 (8)	1.0105+/-0.0000 (2)	0.3647+/-0.0000 (8)	5.4
NSD	0.0093+/-0.0000 (4)	0.4118+/-0.0002 (6)	0.3996+/-0.0000 (6)	1.0107+/-0.0000 (3)	0.3978+/-0.0000 (6)	5.0
DFS	0.0082+/-0.0000 (2)	0.5860+/-0.0000 (7)	0.3717+/-0.0000 (7)	1.0108+/-0.0000 (4)	0.3689+/-0.0000 (7)	5.4
BFS	0.0272+/-0.0000 (7)	0.3330+/-0.0000 (2)	0.4087+/-0.0000 (2)	1.0177+/-0.0000 (7)	0.4085+/-0.0000 (2)	4.0
RW	0.0158+/-0.0001 (6)	0.3514+/-0.0012 (3)	0.4079+/-0.0000 (3)	1.0144+/-0.0000 (6)	0.4071+/-0.0000 (3)	4.2
MHRW	0.0082+/-0.0000 (2)	0.3550+/-0.0014 (4)	0.4054+/-0.0000 (5)	**1.0083+/-0.0000** (1)	0.4013+/-0.0000 (5)	**3.4**
FRW	0.0108+/-0.0000 (5)	0.3569+/-0.0010 (5)	0.4061+/-0.0000 (4)	1.0143+/-0.0000 (5)	0.4047+/-0.0000 (4)	4.6
GFNS	0.0374+/-0.0000 (8)	**0.2485+/-0.0000** (1)	**0.4432+/-0.0000** (1)	1.0193+/-0.0000 (8)	**0.4408+/-0.0000** (1)	3.8

(a) Results for German credit dataset.

	δ-CC ($\nearrow$)	Ddist ($\nearrow$)	Harmonic ($\searrow$)	nCGR ($\nearrow$)	min-σ ($\searrow$)	Average rank
NS	0.1770+/-0.0017 (8)	0.6586+/-0.0013 (8)	0.0150+/-0.0000 (8)	1.0390+/-0.0002 (8)	0.0145+/-0.0000 (8)	8.0
NSD	0.0766+/-0.0001 (4)	0.2243+/-0.0007 (3)	0.0794+/-0.0002 (6)	1.0183+/-0.0000 (6)	0.0779+/-0.0002 (6)	5.0
DFS	0.0196+/-0.0000 (2)	0.5384+/-0.0000 (7)	0.0238+/-0.0000 (7)	1.0248+/-0.0000 (7)	0.0233+/-0.0000 (7)	6.0
BFS	0.1050+/-0.0000 (7)	0.3077+/-0.0000 (4)	0.1094+/-0.0000 (3)	**1.0072+/-0.0000** (1)	0.1093+/-0.0000 (3)	3.6
RW	0.1029+/-0.0001 (6)	0.3388+/-0.0008 (5)	0.0933+/-0.0000 (4)	1.0117+/-0.0000 (3)	0.0918+/-0.0001 (4)	4.4
MHRW	0.0406+/-0.0004 (3)	0.1919+/-0.0050 (2)	**0.1455+/-0.0002** (1)	1.0130+/-0.0000 (5)	**0.1444+/-0.0002** (1)	2.4
FRW	0.0957+/-0.0003 (5)	0.3412+/-0.0024 (6)	0.0893+/-0.0001 (5)	1.0125+/-0.0000 (4)	0.0874+/-0.0002 (5)	5.0
GFNS	**0.0143+/-0.0000** (1)	**0.1824+/-0.0000** (1)	0.1353+/-0.0000 (2)	1.0087+/-0.0000 (2)	0.1351+/-0.0000 (2)	**1.6**

(b) Results for Facebook dataset.

	δ-CC ($\nearrow$)	Ddist ($\nearrow$)	Harmonic ($\searrow$)	nCGR ($\nearrow$)	min-σ ($\searrow$)	Average rank
NS	0.5235+/-0.0060 (8)	0.9989+/-0.0000 (8)	0.0004+/-0.0000 (8)	1.0124+/-0.0003 (6)	0.0003+/-0.0000 (8)	7.6
NSD	0.2419+/-0.0074 (7)	0.9970+/-0.0000 (7)	0.0012+/-0.0000 (7)	**1.0000+/-0.0000** (1)	0.0004+/-0.0000 (7)	5.8
DFS	0.1463+/-0.0000 (6)	0.9844+/-0.0000 (6)	0.0061+/-0.0000 (6)	**1.0000+/-0.0000** (1)	0.0057+/-0.0000 (6)	5.0
BFS	0.1207+/-0.0000 (4)	0.8677+/-0.0000 (5)	0.0084+/-0.0000 (3)	1.0506+/-0.0000 (8)	0.0067+/-0.0000 (3)	4.6
RW	**0.0192+/-0.0003** (1)	0.8082+/-0.0020 (4)	0.0074+/-0.0000 (4)	1.0175+/-0.0004 (7)	0.0067+/-0.0000 (3)	3.8
MHRW	0.0806+/-0.0024 (3)	0.5522+/-0.0182 (2)	0.0159+/-0.0000 (2)	1.0014+/-0.0000 (4)	0.0125+/-0.0000 (2)	2.6
FRW	0.0335+/-0.0003 (2)	0.7682+/-0.0016 (3)	0.0069+/-0.0000 (5)	1.0042+/-0.0000 (5)	0.0062+/-0.0000 (5)	4.0
GFNS	0.1451+/-0.0000 (5)	**0.1821+/-0.0000** (1)	**0.0262+/-0.0000** (1)	**1.0000+/-0.0000** (1)	**0.0150+/-0.0000** (1)	**1.8**

(c) Results for Credit default dataset.

	δ-CC ($\nearrow$)	Ddist ($\nearrow$)	Harmonic ($\searrow$)	nCGR ($\nearrow$)	min-σ ($\searrow$)	Average rank
NS	**0.0005+/-0.0000** (1)	0.3981+/-0.0000 (8)	**0.2169+/-0.1182** (1)	2.0519+/-0.0000 (8)	**0.2077+/-0.1155** (1)	3.8
NSD	**0.0005+/-0.0000** (1)	0.1051+/-0.0034 (2)	0.0122+/-0.0008 (8)	1.9512+/-0.0000 (7)	0.0096+/-0.0005 (8)	5.2
DFS	**0.0005+/-0.0000** (1)	0.2981+/-0.0000 (6)	0.0262+/-0.0000 (6)	1.0427+/-0.0000 (2)	0.1069+/-0.0000 (4)	3.8
BFS	**0.0005+/-0.0000** (1)	0.1931+/-0.0000 (3)	0.1147+/-0.0000 (2)	1.0721+/-0.0000 (6)	0.1131+/-0.0000 (2)	**2.8**
RW	**0.0005+/-0.0000** (1)	0.2998+/-0.0002 (7)	0.1142+/-0.0000 (3)	1.0530+/-0.0000 (4)	0.1059+/-0.0000 (5)	4.0
MHRW	0.0018+/-0.0000 (8)	0.2668+/-0.0052 (4)	0.0153+/-0.0001 (7)	1.0673+/-0.0002 (5)	0.0124+/-0.0001 (7)	6.2
FRW	**0.0005+/-0.0000** (1)	0.2806+/-0.0000 (5)	0.1142+/-0.0000 (3)	1.0519+/-0.0000 (3)	0.1110+/-0.0000 (3)	3.0
GFNS	**0.0005+/-0.0000** (1)	**0.0821+/-0.0000** (1)	0.0786+/-0.0000 (5)	**1.0402+/-0.0000** (1)	0.0682+/-0.0000 (6)	**2.8**

(d) Results for Tagged dataset.

without transitioning to its neighbor [19]. MHRW converges to a uniform distribution unlike RW, whose stationary distribution is proportional to the node degree.

B. Experimental Results

1) General Performance: For fair comparison, all methods are initialized with the same 10 randomly selected target nodes. Table II shows the results on the four datasets. Based on their overall rankings, GFNS outperforms all other methods in 3 of the 4 datasets and is the second best method for the German credit dataset. The MHRW approach is the next best approach, achieving the best rating for the German credit dataset and second best for the Facebook and Credit default datasets. However, it performs poorly on the Tagged dataset.

In terms of structural preservability (δ-CC, Ddist, and Harmonic), GFNS performs relatively better than other baselines. It consistently appears among the top-2 best methods in 9 out of 12 settings (4 datasets and 3 evaluation metrics). For the Tagged dataset, all the methods except for MHRW were able to achieve a small value for δ-CC as the graph is large and highly sparse. As shown in Table I, the average clustering coefficient for the Tagged data is much smaller than other datasets.

In terms of the group representativity metrics such as nCGR and min-σ, the results in Table II suggest that GFNS generally performs better than other baseline methods, including the fair random walk (FRW) method [18]. In fact, GFNS appears among the top-2 best methods in 6 out of the 8 settings (4 datasets and 2 evaluation metrics). It has the best performance in terms of both nCGR and min-σ criteria on the Credit default

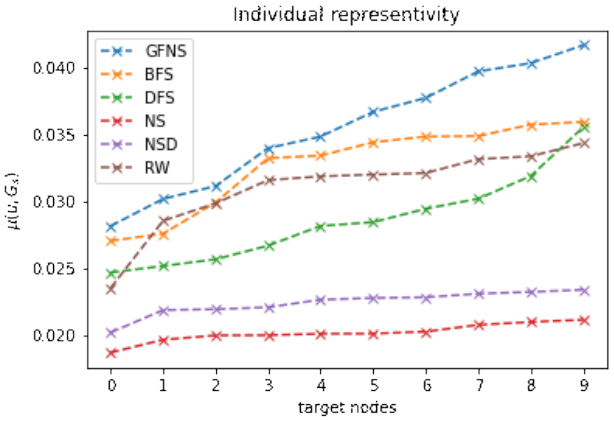

Fig. 1: Comparison of harmonic centrality ratio values of 10 target nodes in the Facebook dataset.

dataset. For the rest of the datasets, GFNS always achieves the best performance for at least one of the two evaluation metrics. These results validated the advantages of using GFNS as a fair graph sampling algorithm that produces unbiased samples while preserving the topological properties of the network.

2) Performance on Target Nodes: We also evaluate the performance of the sampling algorithms in terms of preserving the harmonic centrality of the target nodes. Figure 1 shows the centrality ratio value (μ) for 10 selected target nodes with the highest harmonic centrality from the Facebook dataset. We ran each algorithm to create a sample that contains 100 (2.5%) nodes in the original network. We then sorted the μ values of the target nodes and plot their sorted values in Figure 1. As expected, GFNS outperforms all the baseline methods in terms of preserving the harmonic centrality of the target nodes. The plot shows that the harmonic centrality ratio is consistently higher for GFNS for all 10 target nodes. BFS is the next best performing algorithm, followed by RW.

Finally, we examine the average harmonic centrality of the 10 target nodes when sample size is varied from 100 to 400. Table III summarizes the average centrality ratio of the target nodes for the German credit dataset. A higher value of average centrality ratio indicates better performance in terms of preserving structural properties of the target nodes. For this experiment, we randomly select 10 nodes, five from each gender, as target nodes and repeat the experiment 10 times. The mean and standard deviation of the average centrality ratio of the target nodes are shown in Table III. The result suggests that GFNS outperforms other methods in preserving harmonic centrality of the target nodes irrespective of the sample size.

TABLE III: Comparison of average harmonic centrality ratio values for target nodes in the German credit dataset as sample size is varied.

	100	200	400
NS	0.0617+/-0.0001	0.1629+/-0.0001	0.3639+/-0.0001
NSD	0.0841+/-0.0000	0.1900+/-0.0000	0.3984+/-0.0000
DFS	0.0765+/-0.0000	0.1651+/-0.0000	0.3672+/-0.0000
BFS	0.0985+/-0.0000	0.2033+/-0.0000	0.4211+/-0.0000
RW	0.0850+/-0.0000	0.2013+/-0.0000	0.4129+/-0.0000
FRW	0.0870+/-0.0000	0.1984+/-0.0000	0.4114+/-0.0000
GFNS	**0.1147+/-0.0000**	**0.2259+/-0.0000**	**0.4396+/-0.0000**

V. Conclusions

This paper presents a novel fairness-aware network sampling approach that combines the structural preservability and group representativity objectives into a unified learning framework. We introduced a new sugraph fairness criterion and developed a greedy fair network sampling algorithm with well-grounded theoretical bounds on the greedy approximation. Finally, we experimentally demonstrate the effectiveness of the proposed method on various real-world network data.

VI. Acknowledgement

This material is based upon work supported by the NSF Program on Fairness in AI in collaboration with Amazon under grant IIS-1939368. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or Amazon.

References

[1] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *KDD*, 2006, pp. 631–636.
[2] W. Wei, J. Erenrich, and B. Selman, "Towards efficient sampling: Exploiting random walk strategies," in *AAAI*, vol. 4, 2004, pp. 670–676.
[3] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *Proc. of the 10th ACM SIGCOMM Conf. on Internet Measurement*, 2010, pp. 390–403.
[4] O. Frank and T. Snijders, "Estimating the size of hidden populations using snowball sampling," *Journal of Official Statistics-Stockholm-*, vol. 10, pp. 53–53, 1994.
[5] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *KDD*, 2005, pp. 177–187.
[6] P. Hu and W. C. Lau, "A survey and taxonomy of graph sampling," *arXiv preprint arXiv:1308.5865*, 2013.
[7] C. Wagner, P. Singer, F. Karimi, J. Pfeffer, and M. Strohmaier, "Sampling from social networks with attributes," in *WWW*, 2017, pp. 1181–1190.
[8] N. Martinez, M. Bertran, and G. Sapiro, "Minimax Pareto fairness: A multi objective perspective," in *ICML*, 2020, pp. 6755–6764.
[9] M. Newman, *Networks: An introduction*. Oxford university press, 2010.
[10] P. Boldi and S. Vigna, "Axioms for centrality," *Internet Mathematics*, vol. 10, no. 3-4, pp. 222–262, 2014.
[11] P. Boldi, A. Luongo, and S. Vigna, "Rank monotonicity in centrality measures," *Network Science*, vol. 5, no. 4, pp. 529–550, 2017.
[12] A. A. Bian, J. M. Buhmann, A. Krause, and S. Tschiatschek, "Guarantees for greedy maximization of non-submodular functions with applications," in *ICML*, 2017, pp. 498–507.
[13] A. Das and D. Kempe, "Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection," in *ICML*, 2011, pp. 1057–1064.
[14] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *NeurIPS*, 2012, pp. 539–547.
[15] S. Fakhraei, J. Foulds, M. Shashanka, and L. Getoor, "Collective spammer detection in evolving multi-relational social networks," in *KDD*, 2015, pp. 1769–1778.
[16] D. Dua and C. Graff, "UCI Machine Learning Repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml
[17] I.-C. Yeh and C.-H. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Exp. Sys. with Appl.*, vol. 36, no. 2, pp. 2473–2480, 2009.
[18] T. A. Rahman, B. Surma, M. Backes, and Y. Zhang, "Fairwalk: Towards fair graph embedding." in *IJCAI*, 2019, pp. 3289–3295.
[19] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger, "On unbiased sampling for unstructured peer-to-peer networks," *IEEE/ACM Transactions on Networking*, vol. 17, no. 2, pp. 377–390, 2008.